Behaviour Research and Therapy 192 (2025) 104787

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

# Induced awareness of stimulus differences impedes the generalization of extinction learning with various, highly similar stimuli

Paula Engelke [a,*], Alex H.K. Wong [b], Andre Pittig [c]

[a] Clinical Psychology and Psychotherapy, Institute of Psychology, University of Würzburg, Würzburg, Germany
[b] Department of Psychology, Education, and Child studies, Erasmus University Rotterdam, Rotterdam, Netherlands
[c] Translational Psychotherapy, Institute of Psychology, University of Göttingen, Göttingen, Germany

ARTICLE INFO

*Keywords:*
Exposure
Anxiety disorders
Extinction
Generalization
Perceptual discrimination
Stimulus variability
Awareness of perceptual differences

ABSTRACT

Fear extinction procedures serve as a laboratory model for a learning process involved in exposure treatment for anxiety disorders. Clinically, exposure is typically conducted with generalization stimuli (GSs) because originally acquired fear stimuli are inaccessible. Experimental studies, however, show limited generalization of extinction when GSs are used in extinction training (i.e., GS extinction). Stimulus variability may overcome limited extinction generalization. In a first study, we examined whether using multiple, perceptually similar GSs during fear extinction may overcome limited GS extinction generalization. A healthy sample (N = 120) underwent a two-day fear conditioning procedure, with three groups receiving extinction training with a single GS, multiple GSs or the original CS+. Surprisingly, no group differences appeared. Missing awareness of stimuli differences may explain these findings. Hence, in a second study, we aimed to manipulate awareness of stimulus differences between two groups (N = 80), using additional instructions and tasks. Results suggest that enhanced stimulus differentiation was successfully induced, and pivotally influenced extinction learning and generalization. Specifically, greater awareness of stimulus differences resulted in limited GS extinction generalization toward the original CS+. Results highlight a pivotal role of stimulus differentiation, presumably in interaction with attentional and higher-order cognitive processes, for GS extinction learning and its generalization.

## 1. General introduction

Exposure-based interventions are a crucial part of cognitive behavioral therapy (CBT), the psychotherapeutic first-line treatment for anxiety disorders (Craske et al., 2018; Pittig et al., 2018). Although CBT is known to be highly effective (Carpenter et al., 2018; Hoyer et al., 2017; van Dis et al., 2020), not all patients benefit equally well and fast, and some may experience a return of fear (Boschen et al., 2009; Pittig et al., 2021). A better understanding of the underlying mechanisms of change may boost efficacy. In this regard, CBT is based on multiple mechanisms of change (Benito et al., 2024). A prominent mechanism is the learning that an anticipated threat does not occur during exposure (Craske et al., 2014; Pittig et al., 2023). Fear extinction training serves as one important laboratory model to investigate learning during exposure (Scheveneels et al., 2016). Here, the repeated confrontation with a conditioned fear stimulus (CS+) in the absence of a threatening outcome (i.e., extinction training) will lead to a decrease of conditioned fear. This procedure is assumed to model the confrontation with an anxiety-provoking stimulus or situation during exposure. Expectancy change is thought to be a core mechanism of extinction learning, which may follow expectancy violation (Craske et al., 2018; Pittig et al., 2016; Rescorla & Wagner, 1972). A higher mismatch of the expectancy, that a stimulus will be followed by a threatening event will therefore result in stronger learning.

Typically, laboratory extinction training is conducted with the so-called “original CS+”, i.e., the CS to which conditioned fear was initially acquired. As the CS+ is linked to high threat expectancy, fear extinction with the original CS+ triggers a strong expectancy violation. However, the clinical analogue of the “original CS+” is typically not accessible in exposure treatment, as it is unknown or cannot be reproduced accurately. Thus, exposure exercises are often performed with a perceptually or semantically related stimulus or situation. In the laboratory, the use of such generalization stimuli (GSs) may be less effective compared to the original CS+, leading to stronger return of fear (ROF) later on (e.g., Vervliet et al., 2004, 2005; Vervoort et al., 2014; Wong & Lovibond, 2020). Recent clinical data also provide tentative evidence for

* Corresponding author.
*E-mail address:* paula.engelke@uni-wuerzburg.de (P. Engelke).

https://doi.org/10.1016/j.brat.2025.104787
Received 22 October 2024; Received in revised form 13 April 2025; Accepted 27 May 2025
Available online 28 May 2025

limited extinction generalization. Across multiple exposures, threat expectancy before exposure only slightly declined across sessions (Pittig et al., 2023). Given that exposure situations and stimuli were mostly varied throughout exercises, the amount of generalization from one to the next session seems to be limited. Therefore, research on underlying mechanisms as well as strategies to boost extinction generalization could help to inform treatment and improve long-term learning from exposure.

For developing strategies to improve (limited) extinction generalization, the understanding of influencing factors seems warranted. Following inhibitory learning models (e.g., Bouton, 1993), fear extinction can be explained with a change in CS+ meaning: While an excitatory association between the elements of the CS+ and the occurrence of the aversive stimulus (US) is established during fear acquisition, a competing inhibitory CS+-noUS association is formed during CS+ extinction training. Notably, while the proportion and strength of excitatory and inhibitory associations shape fear reactions, both seem to stay intact throughout extinction learning processes (Bouton, 2002). This, in turn, can explain the possibility of a return of fear (ROF) after extinction training, for example as a function of time (*spontaneous recovery*) or triggered via confrontation with the US (*reinstatement;* see Haaker et al., 2014; Vervliet, Craske, et al., 2013). When extinction training is conducted with the original CS+, the presentation of a similar GS does not lead to a strong ROF, suggesting a generalization of CS+-extinction learning towards other similar stimuli (Vervliet et al., 2004; Wong & Lovibond, 2020). In contrast, when extinction training is conducted with a GS, learning seems to generalize only to a limited extent back to the original CS+ or to other related stimuli (Barry et al., 2016; Vervliet et al., 2004, 2005; Wong & Lovibond, 2020). Importantly, a perceptually similar GS consists of elements overlapping with the CS+, which possess excitatory strength after fear acquisition, but also distinct features not shared with the CS+. When presenting a GS in extinction training, the shared elements will gain inhibitory strength, leading to extinction learning (Barry et al., 2017; Vervliet et al., 2005). However, in a later confrontation with the original CS+ or other similar stimuli, the elements of the CS+ not shared with the extinguished GS still maintain excitatory strength and therefore elicit a strong conditioned fear. In support of the suggested role of similarity (i.e., overlap of elements), extinction learning was found to be a function of perceptual similarity of the GS in extinction training, with fewer similarity (i.e., fewer shared, extinguished elements) triggering stronger return of US expectancy (Vervliet et al., 2006). Limited GS extinction generalization was shown for perceptually but also conceptually or categorically related stimuli (e.g., Gerdes et al., 2020; Vervoort et al., 2014; Zbozinek & Craske, 2018). Notably, however, GS extinction learning does not *necessarily* lead to limited generalization under certain conditions: For example, the presentation of a peak GS (i.e., representing a more intense version of the CS+) was shown to even outperform CS+ extinction learning (Struyf et al., 2018). Hence, the investigation of further ways to enhance the generalization of GS extinction learning appears promising.

In a first experiment, we aimed to test the use of multiple, perceptually similar GSs in contrast to one single GS as a potential strategy to improve GS extinction generalization. Therefore, 120 healthy participants underwent a two-day fear conditioning procedure. In a second experiment, we further investigated the influence of the awareness of stimulus differences on (GS) extinction generalization. Hence, another 80 participants underwent an adapted version of the first experiment.

## 2. First experiment

### 2.1. Specific introduction

In the first experiment, we aimed to use multiple, perceptually varying GSs in contrast to one single GS as a strategy to improve generalization of GS extinction learning (Craske et al., 2018; Lipp et al., 2020). Since each of the GSs shares features with the CS+ but also differs from each other in their distinct characteristics, the buildup of a more general inhibitory association and its retrieval could be facilitated (Craske et al., 2018, 2022). Empirical studies with clinical-analog (i.e., participants with phobia) and healthy samples found mixed evidence on whether extinction training with variable GSs can improve extinction generalization. While some analog studies indicated that presenting multiple GSs during exposure could reduce ROF or enhance approach behavior (Rowe & Craske, 1998; Shiban et al., 2015), others found null evidence (Kircanski et al., 2012; Lang & Craske, 2000). Likewise, experimental studies found a positive effect for presenting multiple GSs in addition to the CS+ in extinction training (Waters et al., 2018, 2021). In contrast, the use of multiple, categorically related GSs did not improve the generalization of extinction learning back to the CS+ or a novel stimulus (Zbozinek & Craske, 2018). Notably, in the forementioned studies, either conceptually associated GSs were investigated, or the level of perceptual similarity between stimuli could not be controlled (i.e., in clinical-analog studies). Although perceptual (dis) similarity seems to be one driving factor for limited extinction generalization (e.g., Vervliet et al., 2006), the effect of multiple *perceptually* similar GSs in extinction training remains unclear. Therefore, we used several morphed face stimuli with a fixed level of similarity to the CS+ for extinction training with multiple different GSs.

To test stimulus variability as a strategy to improve extinction generalization, either one single GS, multiple GSs or the original CS+ served as extinction stimuli in three randomized groups. After fear acquisition and extinction training, a yet unknown facial stimulus (novel stimulus, NS) and the original CS+ were presented in an immediate test phase in all groups. Furthermore, retention and reinstatement tests were applied after 24 h. We hypothesize that extinction training with a single GS will be associated with a return of conditioned fear across test phases, i.e., stronger fear reaction to CS+ and novel stimuli (NS), compared to extinction training with the CS+. In contrast, extinction training with multiple GSs is assumed to be associated with less return of fear than single GS extinction training (i.e., strategy of stimulus variability). Multiple GSs vs. CS+ extinction training will be compared exploratorily. Furthermore, potential discrepancies in extinction learning slopes due to different extinction training procedures between groups will be examined. Finally, we hypothesize that fear reaction to the CS+ should increase across test phases, considering the elapse of time (spontaneous recovery) and reinstatement manipulation. However, the course of fear responses to the NS across test phases will be analyzed exploratorily. Aiming to verify laboratory findings, 120 additional participants underwent a web-based version of the paradigm. As the results aligned with laboratory findings, specific description of method and results of the web-based paradigm were included in the Supplementary Information (see SI, S2.5).

### 2.2. Method

The study design and sample size estimation for the first experiment were preregistered (https://osf.io/fdtvm) and procedures were approved by the ethics committee of the Psychological Institute of the University of Würzburg (GZ 2020–29) in accordance with the Declaration of Helsinki.

#### 2.2.1. Participants

Aiming at a sample size of 40 participants per group, 121 healthy participants between 18 and 50 years were recruited from the general community and students at the University of Würzburg. Exclusion criteria included self-reported cardiovascular, respiratory, neurological, or hearing disorders, diagnosed emotional or affective disorders as well as any other serious health conditions, self-reported substance dependence, use of psychotropic medication, traumatic brain injuries, and pregnancy. One participant declined participation after one presentation of the loud scream (later US) was provided exemplarily (i.e., before giving written consent). Therefore, 120 participants were randomly assigned to one of three groups and tested by the same researcher on both study days. Four participants (4.2 %) did not attend the second day of the study. In this case, as preregistered, data from the first day were kept for analysis. Missing questionnaire items were replaced using single imputation (i.e., participants scale mean). Groups did not differ regarding age, sex, trait anxiety, depression symptoms, anxiety symptoms, and intolerance of uncertainty (see Table 1).

### 2.2.2. Stimuli

All face stimuli used during conditioning procedure are part of the Karolinska Directed Emotional Faces (*KDEF*) set (Lundqvist et al., 2015). One neutral female face stimulus served as the designated conditioned stimulus (CS+), while the aversive unconditioned stimulus (US) consisted of a 95 dB scream (International Affective Digital Sounds System) and fearful-faced picture of the CS+, presented for 1.5 s (e.g., Lau et al., 2008). Morph-stimuli were prepared using the software FantaMorph (Version 5.0, Abrosoft Co.): The CS+ was individually morphed with one of six other stimuli from the same stimulus set, resulting in generalization stimuli (GSs) with graded similarity levels (e.g., 70 % to the CS+ and 30 % to another stimulus). Prior to data collection, the distinguishability of the morphed stimuli was validated in two web-based studies (see SI, S1). Mimicking the main task, participants classified the sequentially presented stimulus pictures as identical or different. Results indicated that GSs resembling the CS+ face to 70 % could be discriminated safely from the original CS+ (overall hit rate: 82.33 %, $p < .001$ in binomial test). In contrast, discrimination for 80 % GSs was at chance (54.00 %, $p = .184$), and 90 % GSs were often misinterpreted being the CS+ (21.00 %, $p < .001$). Additionally, 70 % GSs could be safely distinguished from each other (71.73 %, all $ps \leq .004$), and were thus selected for the main task. Three neutral female face stimuli were selected as novel stimuli (NS1 to 3). Of note, these stimuli were different than the ones used for GSs preparation. Therefore, they represent novel stimuli that share overall characteristics with the CS+ (i.e., female face) but, unlike GSs, are not thought to share specific facial features.

### 2.2.3. Procedure

Due to the COVID-19 pandemic, hand disinfection in the beginning and wearing of masks were mandatory and participation required a valid vaccination or test status in accordance with public health measures. Participants provided written consent after hearing one example of the loud noise (i.e., scream, 95 dB) over earphones, which was later used as the US. After that, participants filled out a battery of questionnaire battery assessing trait anxiety (anxiety facet of the NEO-PI R – N1, Costa & McCrae, 1992), anxiety symptoms (PROMIS Short Form v1.0 - Anxiety 8a; Wahl et al., 2011), symptoms of depression (ADS-L, Hautzinger et al., 2012), and intolerance of uncertainty (UI-18, Gerlach et al., 2008). Half of the questionnaires were filled out at the second day. Next, on either day, two reusable Ag/AgCl electrodes were at participants' non-dominant hand for skin conductance recording. Then, all participants received a brief introduction and went through a fear conditioning paradigm (see Table 2). Notably, fear conditioning experiments often introduce both a CS+ and a safe CS- (i.e., differential conditioning). However, if a CS- is part of the same continuum as the CS+ (i.e., size of a geometrical form), disentangling extinction learning mechanisms can be difficult: Both a GS, sharing excitatory features with the CS+, and a CS-, sharing inhibitory features with the GS, are presented in extinction training. Therefore, the decrease of fear can be attributed either to the extinguished excitatory elements, the inhibitory elements of the CS-, or a combination of both (Wong & Lovibond, 2020). To unravel the sole effect of extinguishing shared excitatory elements, only a CS+ was trained in the first experiment (i.e., single-cue conditioning).

The experiment was programmed using OpenSesame Software (Version 3.3, Mathôt et al., 2012). On Day 1, a female face picture (designated CS+) was presented four times in front of a black background without any aversive US (*habituation phase*). In the following *fear acquisition training*, the CS+ was presented ten times, with eight presentations immediately followed by the aversive US. The trial order was pseudorandomized, with the first and last acquisition trial being always reinforced. Next, participants underwent one of three different *extinction trainings* without US presentation: In the Single GS group, one 70 % GS was presented 12 times. In the Multiple GSs group, six different 70 % GSs were presented randomized across two blocks, while the CS+ Control group experienced 12 presentations of the original CS+. Subsequently, all participants underwent an *immediate test phase* with the presentation of one novel stimulus (NS1, different female face) followed by one presentation of the original CS+ without reinforcement. Of note, this fixed sequence was implemented to control for unbalanced extinction effects between groups. After 24 h, the participants returned to the laboratory. Here, one trial with another novel face (NS2) followed by one trial of the CS+ were presented (*retention test phase*). Next, to induce reinstatement of fear, two unsignaled, auditory USs were presented (i.e., scream, ISI 15000–18000 ms). Afterward, another test phase, including a third novel stimulus (NS3) and the CS+ (*reinstatement test phase*), was conducted. Notably, the assignment of the three novel stimuli to the test phases was counterbalanced.

Participants were instructed in the beginning, that some of the presented pictures might be followed by a loud noise and they have the task to rate the likeliness for that. In each trial, the CS+ or GS stimulus was presented for 8 s in a size of 140×190 mm. Simultaneously, the participants used a scale below to evaluate the likeliness of a loud noise (online US expectancy ratings, 0–100%) via mouse click. Ratings could be adjusted until stimulus offset. A white fixation cross was shown for a duration randomized between 15000 and 18000 ms between trials (ITI). All phases of the experiment (on each day) followed directly after each other without interruption or additional instructions. After the reinstatement test at the Day 2, participants indicated valence and arousal ratings for the CS+ and the recalled US (discrete scales from 1 to 9), and if they would choose to avoid the US (0–100%). Additionally, participants indicated the number of different faces they saw on both days of the study (open question).

**Table 1**
Demographic and Questionnaire Data of the Laboratory Sample in Experiment 1.

| | Experimental Group | | | Statistics | | |
|---|---|---|---|---|---|---|
| | Single GS(n=40) | Multiple GSs (n=40) | CS+ (n=40) | F or $\chi^2$ | $p$ | $\eta^2$ or $r$ |
| Age | 25.38 (7.11) | 25.58 (4.81) | 24.92 (5.16) | 0.13[a] | .876 | .002 |
| Sex (female) | 30 | 28 | 31 | 0.61[c] | .738 | .071 |
| Trait Anxiety (NEO-PI R-N1) | 17.95 (5.86) | 18.38 (5.47) | 18.03 (5.32) | 0.07[b] | .934 | .001 |
| Anxiety Symptoms (PROMIS) | 16.02 (6.84) | 15.45 (4.43) | 14.97 (4.16) | 0.40[a] | .674 | .007 |
| Depression (ADS-L) | 11.38 (9.97) | 11.03 (7.13) | 11.35 (7.73) | 0.02[a] | .979 | <.001 |
| Intolerance of Uncertainty (UI-18) | 42.97 (13.85) | 42.90 (12.39) | 42.11 (13.35) | 0.05[b] | .953 | .001 |

*Note.* Means (and standard deviations) for demographic and questionnaire data per group in Experiment 1; NEO-PI-R-N1: anxiety subscale of NEO-PI R, range = 0–32 (Costa & McCrae, 1992); PROMIS: PROMIS Short Form v1.0 - Anxiety 8a, range = 8–40 (Wahl et al., 2011); ADS-L = General depression scale – Long form, range 0–60 (Hautzinger et al., 2012); UI-18: Intolerance of Uncertainty Scale, range 18–90 (Gerlach et al., 2008); [a]$F(2,117)$; [b]$F(2,113)$; [c] $\chi^2(2)$; differences in degrees of freedom are due to the dropout of four participants at second test day.

**Table 2**
Experimental Design of Experiment 1.

| Group | Day 1 | | | | | Day 2 | |
|---|---|---|---|---|---|---|---|
| | Habi-tuation | Fear Acquisition | Extinction Training | Immediate Test | Retention Test | Rein- statement | Reinst. Test |
| Single GS | CS+ (4) | CS+$_{80\%}$ (10) | GS1 (12) | NS1 (1)<br>CS+ (1) | NS2 (1)<br>CS+ (1) | US (2) | NS3 (1)<br>CS+ (1) |
| Multiple GSs | CS+ (4) | CS+$_{80\%}$ (10) | [GS1-GS6] (2) | NS1 (1)<br>CS+ (1) | NS2 (1)<br>CS+ (1) | US (2) | NS3 (1)<br>CS+ (1) |
| CS+ | CS+ (4) | CS+$_{80\%}$ (10) | CS+ (12) | NS1 (1)<br>CS+ (1) | NS2 (1)<br>CS+ (1) | US (2) | NS3 (1)<br>CS+ (1) |

*Note.* CS+ = original face stimulus; during Fear Acquisition, the CS+ was reinforced by the aversive US (scream and fearful face) in 80 % of trials; the CS+ was not reinforced in any other phases; GS = generalization stimulus (each with 70 % similarity to CS); NS = novel face stimulus (neutral); US = auditory US (scream); numbers of trials are indicated in parentheses. All face stimuli are part of the KDEF Set (Lundqvist et al., 2015). The following stimuli were used: AF19NES (CS+), AF19AFS (US); for the preparation of morph stimuli, the following stimuli were each individually morphed with the CS+: AF01NES, AF05NES, AF07NES, AF09NES, AF22NES, AF24NES; novel face stimuli (NS1-3): AF29NES, AF30NES, BF11NES.

#### 2.2.4. *Data processing and analyses*

Throughout the procedure (32 trials), online US expectancy ratings were assessed, and skin conductance was measured. Detailed information about processing of skin conductance data and results are included in the SI, S2.3. Data and analyses are available on https://osf.io/qk2r6/. Analyses were conducted using *R* 4.3.1 (R Core Team, 2023) with the *afex* package (v1.3.0, Singmann et al., 2023). As preregistered, separate linear mixed models (LMMs) for each dependent variable (US expectancy, SCR) and phase of the experiment were computed (see below). Continuous predictors were centered and scaled ($M = 0$, $SD = 1$), while *sum*-to-zero *contrasts* were used for categorical predictors. In addition to main effects-only models (i.e., without interaction terms), models including interaction effects between factors were computed (Hayes et al., 2012). Since the models yielded the same significant main effects, the estimates from interaction models are reported. To evaluate effects, type-III-like p-values using the Satterthwhaites approximation for degrees of freedom were obtained. A standard inference criterion of $p < .05$ was used. For follow-up analyses, estimated marginal means (EMMs) or estimated linear trends (slope estimates) were compared using the *emmeans* package (v1.8.9, Lenth, 2023), with Bonferroni-Holm correction for multiple comparisons. Assumptions of LMMs (i.e., normal distribution of residuals, homoscedasticity) were checked using the *DHARMa* package (v0.4.6, Hartig, 2022). If violations were noticed, follow-up robust linear mixed models with the package *robustlmm* (v3.2.3, Koller, 2016) were computed additionally (see also SI, S2.2).

*2.2.4.1. Fear learning and extinction training models.* Separate models for the habituation, acquisition, and extinction training phases were computed. *Group* and *Trial Number* were included as fixed factors. To model learning trends throughout trials, Trial Number thereby served as a continuous linear predictor. *Participants* served as random factor, including both a random intercept and a random slope for each participant across Trial Number.

*2.2.4.2. Test phase models.* Of note, models for separately analyzing both types of test stimuli (i.e., NS, CS+) were preregistered. To be able to a compare the fear reaction towards both stimuli as well, NS and CS+ were however included within one model: *Stimulus* (CS+ vs. Novel Stimulus NS), *Test Phase* (Immediate vs. Retention vs. Reinstatement Test) *and Group* served as fixed factors, and a random intercept for *Participants* was included. As all test phases only included a single trial of each stimulus (CS+ and NS), no trial number factor was included. Preregistered separate models for NS and CS+ revealed consistent results. To examine the transition between extinction training and immediate test, cross-phase analyses including the last extinction trial and the immediate test phase on Day 1 were computed. Here, *Group* and *Trial* (last extinction trial vs. NS test vs. CS+ test; categorical predictor) were stated as fixed factors with *Participants* serving as random factor.

*2.2.4.3. Exploratory analyses.* Participants were asked in the end of the second day to retrospectively state how many different faces they had seen on the first day. This rating was compared between groups. Furthermore, robust SCR analyses pointed towards group differences during fear acquisition training, despite all groups undergoing the same procedure. Hence, covariate models to account for a-priori differences were applied, and two-part mixed effect models (hurdle models) were computed exploratorily (see SI, S2.4).

### 2.3. *Results*

#### 2.3.1. *Habituation, fear acquisition and fear extinction learning*

Results of habituation, fear acquisition, and extinction training are depicted in Fig. 1A. During habituation, US expectancy significantly decreased across trials, $F(1117.00) = 39.44$, $p < .001$ (slope estimate = $-5.22$, $CI_{95}[-6.87; -3.57]$). As expected, there were no significant effects including the Group factor ($ps \geq .796$, see SI, S2.1 for full statistics). Throughout fear acquisition training, the CS+ was paired with the aversive US. Accordingly, expectancy ratings increased over trials, $F(1,117) = 202.30$, $p < .001$, (slope estimate = 10.70, $CI_{95}[9.24; 12.2]$). There was no significant effect including the group factor, supporting a comparable acquisition of conditioned US expectancy across groups ($ps \geq .247$). During extinction training phase, as the US was no longer applied, a decrease in US expectancy ratings was confirmed through a significant main effect of Trial Number, $F(1,117) = 70.63$, $p < .001$ (slope estimate of $-7.88$, $CI_{95}[-9.74; -6.02]$). Again, no significant group effect or interaction appeared ($ps \geq .850$), indicating comparable extinction learning between groups despite different extinction procedures.

#### 2.3.2. *Test phases throughout both days*

Aiming to compare the reaction to the novel facial stimuli (NS) and the CS+ between groups and test phases, an LMM with the factors Group, Test Phase (i.e., immediate test, retention test, reinstatement test) and Stimulus Type (i.e., NS, original CS+) was computed. For US expectancy, contrary to our hypotheses, neither the three-way interaction nor any effect involving Group was significant ($ps \geq .660$). Therefore, there was no evidence for different reaction between groups. However, Test Phase and Stimulus Type showed a significant interaction, $F(2, 571.12) = 17.90$, $p < .001$ (see Fig. 1B). Follow-up contrasts revealed significant US expectancy increase for the CS+ from the immediate test to both retention and reinstatement tests on the second day, $t(574) = -4.71$, $p_{adj} < .001$, $t(574) = -5.91$, $p_{adj} < .001$, with no significant difference between the latter two, $t(571) = -1.19$, $p_{adj} = .702$. In contrast, NS ratings appeared more stable, as no evidence for a difference between the test phases was shown ($ps_{adj} > .209$). Consistently, comparing NS and CS+ ratings, EMMs showed no significant difference during immediate test phase on the first day, $t(571) = .01$, $p_{adj} = 1.000$, but significantly higher ratings for the CS+ in retention, $t(571) = 6.36$, $p_{adj} < .001$, and reinstatement test phases, $t(571) = 7.90$, $p_{adj} < .001$. Overall, the interaction reflects rather increasing US expectancy over

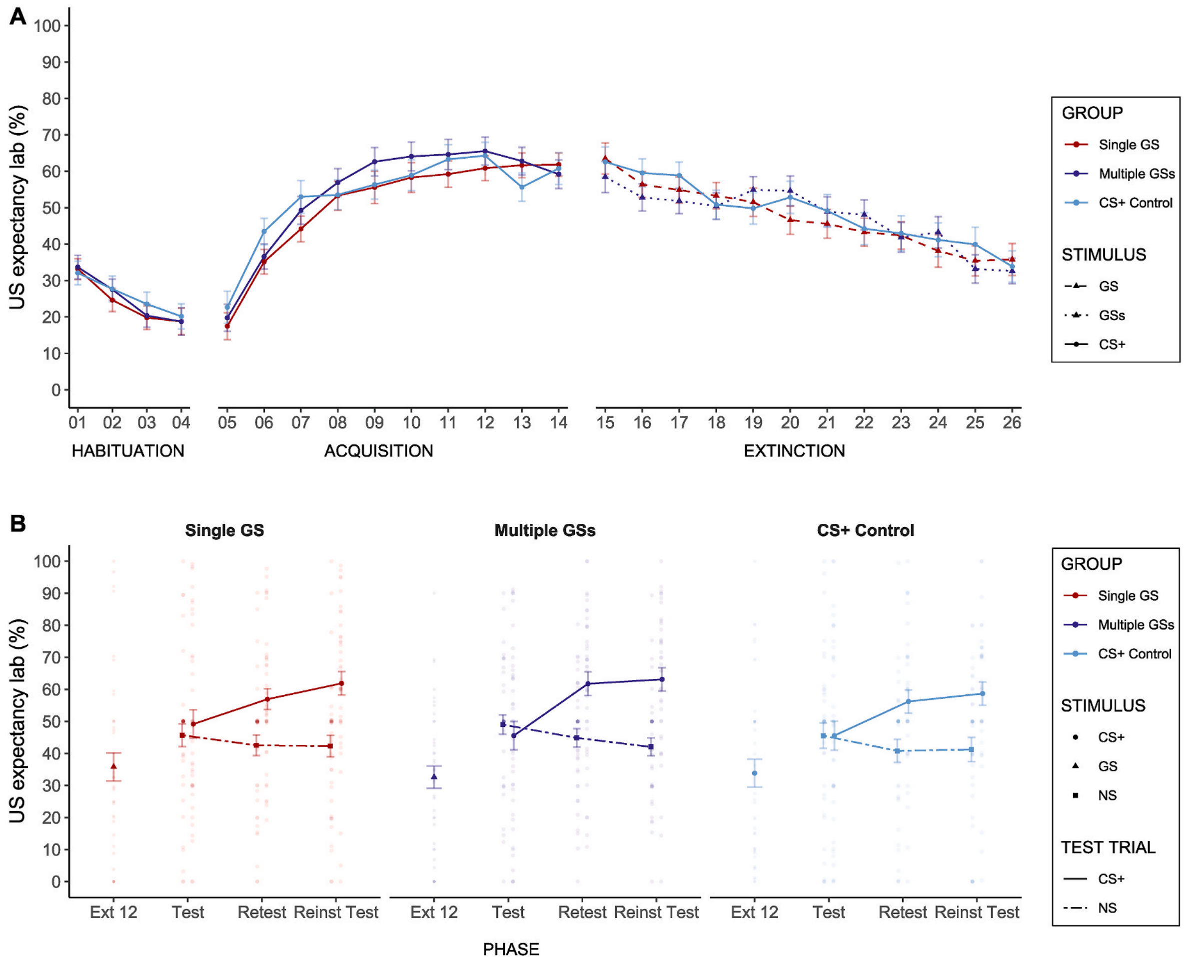


**Fig. 1.** (A) Trial-wise mean US expectancy ratings in the first experiment during habituation, fear acquisition and extinction training phase. During habituation and acquisition training phases, the CS+ was presented in all groups. In extinction training, one GS was presented for the Single GS group, six different GSs for the Multiple GSs group, and the CS+ for the CS+ Control group. Error bars indicate the standard error of the mean. (B) Mean US expectancy ratings per group during the last trial of extinction training (Ext 12) and the three Test Phases: Immediate Test at Day 1 (Test), Retention Test (Retest) and Reinstatement Test (Reinst Test) at Day 2. Notably, the last extinction trial differed between groups: For the Single GS and Multiple GSs groups, a GS was presented, while a CS+ was presented in the CS+ Control group. In all test phases, one NS and the CS+ were presented once (same procedure in all groups). Solid lines represent the CS+, dot-dashed lines represent the novel stimuli (NS1 to 3) across test phases. Error bars indicate the standard error of the mean. Raw participant data are depicted in the background.

test phases for the CS+, while NS ratings remained stable.

To investigate the transition between the end of extinction training and immediate test phase, cross-phase analyses were computed (see Fig. 1B). Here, the three last trials of the first study day (last extinction trial, immediate test with NS and immediate test with CS+) were analyzed. Similar with the previous results, no significant main effect or interaction involving the factor Group for US expectancy was evident (*ps* $\geq .807$). However, a significant Trial effect, $F(2, 234.00) = 16.19$, $p < .001$, indicated increased US expectancy ratings from the last extinction trial to both NS, $t(234) = -4.93$, $p_{adj} < .001$, and CS+ in test, $t(234) = -4.93$, $p_{ad} < .001$. Ratings for the two test stimuli appeared to be not significantly different, $t(234) = .01$, $p_{ad} = .996$.

SCR results are included in the SI, S2.2. While no group difference in habituation phase was observed, robust SCR analyses pointed towards a stronger SCR in the Multiple GSs group during fear acquisition training, despite all groups undergoing the same procedure. Preregistered and exploratory analyses also suggested higher SCR for the Multiple GSs group during extinction training. However, no group differences were found during test phases. Instead, SCR data pointed towards a novelty effect of the NS particularly during immediate test phase, with a strong increase of SCR in all groups. In sum, SCR data did not indicate differences in extinction generalization between groups. The results for the additional web-based sample (n = 120) aligned with laboratory findings for US expectancy ratings: Successful fear acquisition and extinction learning without groups differences was shown. However, learning gradients appeared to be generally less steep. No group differences during fear extinction training or test phases were visible for the web-based sample, replicating the findings of the laboratory task (detailed results in SI, S2.5).

### 2.3.3. *Retrospectively remembered faces*

After the end of the conditioning procedure on Day 2, participants were asked to indicate how many different faces they had seen on Day 1 (open question). A simple-effect Analysis of Variance (ANOVA) revealed a significant group difference, $F(2,113) = 3.38$, $p = .038$, $\eta^2 = .056$. Post-hoc paired and adjusted t-tests showed a significantly higher rating for the Multiple GSs group ($M = 2.59$, $SD = 1.16$) compared to the Control group ($M = 2.11$, $SD = .40$), $t(113) = 2.52$, $p_{adj} = .040$. The Single GS group ($M = 2.25$, $SD = .74$) did not differ from the other groups ($ps_{adj} \geq .146$). However, the Multiple GSs group appeared to underestimate the number of faces strongly (8 faces seen vs. $M_{remembered} = 2.59$), and only 10 participants from that group indicated three or more faces. Thus, GSs were not as successfully differentiated as in the two validation studies.

### 2.4. Discussion of experiment 1

In Experiment 1, we aimed to investigate the generalization of extinction learning after extinction training with the original conditioned fear stimulus (CS+), a single, perceptually similar stimulus (Single GS), or multiple perceptually similar stimuli (Multiple GSs). In both the laboratory and web-based sample, successful fear acquisition and extinction of threat expectancy were shown in all groups. Across three test phases including a novel face stimulus and the original CS+, results indicate an increase of US expectancy for the CS+ from the first to the second test day, but a stable or decreasing response to the novel face stimuli. These findings are in line with spontaneous recovery and reinstatement of conditioned fear to the CS+. However, groups did not differ in their responses across extinction training and test phases. Specifically, we found neither group differences in US expectancy throughout extinction training (i.e., no generalization decrement in the beginning of extinction training), nor an improvement of GS extinction learning when using multiple, similar GSs. The results are surprising as the groups received extinction training with physically different stimuli.

Importantly, our hypotheses presupposed that perceptual differences between the original CS+ and GS(s) can be recognized by participants. Hence, the selection of GSs was based on the aim to use highly similar but distinguishable stimuli. Two prior validation studies provided robust evidence that the 70% GSs can be differentiated from the original CS+ and from each other. From a perceptual point of view, limited GS extinction generalization (i.e., stronger return of fear) can be explained by unextinguished excitatory features of the CS+ that were not "captured" from the GS during extinction training. In a recent study, Waters et al. (2023) compared extinction training with a more versus less similar GS and showed a return of fear only for the less similar GS, while training with the more similar GS led to comparable results as with the CS+. Accordingly, the return of US expectancy was found to be a function of perceptual similarity of the GS extinction stimulus (Vervliet et al., 2006). Due to the high similarity between GS(s) and CS+ in the present study, the GS(s) may have triggered a comparably strong expectancy violation (and therefore extinction learning) as the original CS+. However, as the retrospectively indicated number of different faces only revealed a small effect between groups, there was no strong evidence that participants were able to discriminate between the CS+ and GS(s). Therefore, a lack of awareness of objectively existing differences must be considered as an alternative explanation for the results. That is, misinterpretation of a GS as the CS+ during extinction training could result in an increased conditioned fear reaction also for GSs (Struyf et al., 2017; Zaman, Struyf, et al., 2019), offering an explanation for the observed null group difference during extinction training. In contrast to the validation studies, experiencing stimuli within an aversive context may have impaired perceptual discrimination (Resnik et al., 2011; Schechtman et al., 2010; Zaman et al., 2015). Furthermore, the explicit discrimination task and the slightly shorter ITI could have influenced short-term memory and attentional processes, leading to more accurate discrimination in the validation studies. To further investigate the effect of the awareness of stimulus differences on GS extinction learning and generalization, a second experiment was conducted.

## 3. Second experiment

### 3.1. Specific introduction

Aiming to examine the role of awareness about stimulus differences on GS extinction generalization, 80 novel participants underwent an adapted procedure. Importantly, also 70 % GSs from the first experiment were included as GS extinction stimuli in the second experiment. To manipulate awareness, one group received specific instructions and rated potential stimulus differences in each trial (*Awareness group*). The *Control group*, however, received neutral instructions and rating questions. Verbal instructions are known to be able to influence (GS) extinction learning, for instance via manipulating attention allocation during extinction training (e.g., Barry et al., 2017; Klein et al., 2021; Stemerding et al., 2023; for an overview, see Mertens et al., 2018). Specifically, instructed attention towards unique, distinct features of a GS extinction stimulus (in contrast to emphasizing common characteristics shared with the CS+) resulted in limited generalization of GS extinction learning on physiological level (Barry et al., 2017). While increased attention is thought to enhance visual awareness of differences between non-threat-related stimuli (Kim & Chong, 2022), it has also been shown that a discrimination training for fear-relevant, morphed face stimuli could reduce *fear* generalization towards those stimuli (i.e., enhance differentiation between an original CS+ and GSs, Herzog et al., 2021). However, to the best of our knowledge, *manipulation* of perceptual awareness for differences between highly similar stimuli (i.e., via attention allocation) and its influence on fear extinction learning with GS(s) was not yet investigated. We assume that increased awareness of differences between the original CS+ and similar GSs may promote a return of fear for the CS+ after GS extinction learning (i.e., limited generalization of GS extinction learning).

To be able to directly compare GS and CS+ extinction training *within* participants, two fear stimuli CS+s (i.e., face stimuli) and one safety stimulus CS- from another category (i.e., geometric form) were introduced. During extinction training in both groups, one CS+ was replaced with multiple, highly similar GSs (i.e., see the Multiple GSs group in the first experiment), while the other one remained in its original form. In a subsequent test phase, all original stimuli were presented. We hypothesize that, after successful differential fear acquisition in both groups, the Awareness group will show differential responding between the CS+ and the GSs during extinction training (i.e., generalization decrement in the beginning and overall reduced fear reactions to the GSs). The Control group, however, will not show differential responding between GSs and CS+. Accordingly, we assume an increase of fear after extinction training (i.e., limited GS extinction generalization) *only* for the previously replaced CS+ in the Awareness group, but not for the Control group. Extinction learning and fear reaction during test phase for the non-replaced CS+ and CS- are hypothesized to not differ between both groups. Accompanying ratings (e.g., performance in trial-wise discrimination ratings, retrospective number of faces, memory test) are analyzed for a manipulation check, and potential associations with extinction learning and generalization are explored.

### 3.2. Method

Study design and sample size of the second experiment were preregistered (https://osf.io/yj3kp), and procedures were ethically approved (GZ 2020–29). In sum, 81 participants were recruited with exclusion criteria aligning with the first experiment. Participants from the first experiment were not able to take part. Again, one participant declined participation before giving written consent. 80 participants were randomly assigned to two experimental groups ($n = 40$). Both groups did not differ regarding age, sex, trait anxiety, anxiety, depression or social anxiety symptoms, intolerance of uncertainty, prosopagnosia (i.e., face blindness) traits and their performance in a face matching test (see Table 3).

#### 3.2.1. Stimuli and procedure

Extending the procedure of the first experiment, *two* neutral face stimuli were presented as conditioned fear stimuli (i.e., CS+ A and CS+ B). While one CS+ stimulus was identical with the CS+ used in Experiment 1, the other one was a different neutral female face from the same dataset (Lundqvist et al., 2015). Both CS+ stimuli were counterbalanced between groups. Preparation of six matching 70% GSs for each CS+ face followed the morphing procedure of first experiment (i.e., GSs from Experiment 1 were re-used, and additional ones were prepared; see Table 4). A violet triangle represented the designated CS-. Notably,

**Table 3**
Demographic and Questionnaire Data of Experiment 2.

| | Experimental Group | | Statistics | | |
|---|---|---|---|---|---|
| | Awareness Induction Group (n=40) | Control Group (n=40) | *t* or $\chi^2$ | *p* | d or *r* |
| Age | 23.32 (4.14) | 22.42 (3.14) | $-1.09^a$ | .277 | .245 |
| Sex (female) | 32 | 28 | $0.60^b$ | .439 | .070 |
| Trait Anxiety (NEO-PI R-N1) | 15.75 (3.97) | 14.03 (3.89) | $-1.95^a$ | .054 | .437 |
| Anxiety Symptoms (PROMIS) | 14.55 (4.53) | 15.18 (4.78) | $0.60^a$ | .550 | .134 |
| Depression (ADS-L) | 11.28 (6.82) | 11.82 (6.03) | $0.38^a$ | .703 | .085 |
| Intolerance of Uncertainty (UI-18) | 41.62 (9.19) | 42.83 (10.91) | $0.53^a$ | .596 | .118 |
| Social Phobia Symptoms (SPIN) | 13.28 (8.61) | 13.88 (9.84) | $0.29^a$ | .773 | .065 |
| Prosopagnosia traits (PI20) | 41.70 (10.70) | 43.31 (10.74) | $0.67^a$ | .503 | .150 |
| Performance Face Matching (OFMT) | 149.57 (9.28) | 147.80 (11.45) | $0.75^a$ | .454 | .168 |

*Note.* Means (and standard deviations) for demographic and questionnaire data per group in Experiment 2; NEO-PI-R-N1: anxiety subscale of NEO-PI R, range = 0–32 (Costa & McCrae, 1992); PROMIS: PROMIS Short Form v1.0 - Anxiety 8a, range = 8–40 (Wahl et al., 2011); ADS-L = General depression scale – Long form, range 0–60 (Hautzinger et al., 2012); UI-18: Intolerance of Uncertainty Scale, range 18–90 (Gerlach et al., 2008); SPIN: Social Phobia Inventory, range 0–68 (Connor et al., 2000); PI20 = Prosopagnosia Index, range 20–100 (Shah et al., 2015); OFMT = Oxford Face Matching Test, adapted version, range 0–200 (Stantić et al., 2021, 2024). $^a t(78)$ with Cohens d; $^b\chi^2(1)$ with r.

unlike the first experiment, *both* groups experienced the same stimuli during extinction training, but instructions and trial-wise forced-choice ratings were varied between groups to manipulate awareness of stimulus differences.

Before the conditioning task, participants were instructed they would see different stimuli, and some may be associated with an aversive event. Here, all three original stimuli were depicted below the text. Then, participants rated the three original CS regarding their valence and arousal (continuous scale, −5 to +5). The stimulus presentation size and duration (i.e., 8 s for CSs, 1.5 s for the US), as well as duration of ITI (i.e., randomized between 15 000 and 18 000 ms) aligned with the first experiment. The procedure is depicted in Table 4. Presentation order was pseudo-randomized (i.e., not more than two presentations of a stimulus in a row). In the *habituation phase*, each of the three later CSs was presented two times. Afterward, *fear acquisition training* was carried out in two blocks with five presentations of each CS in each block. In eight out of ten acquisition trials, each CS+ stimulus was directly followed by a matching fearful face and a loud scream (US, see 2.2.2). The CS- was never paired with a US. After the first acquisition block, participants repeated affective ratings and were instructed about an additional forced choice question: To induce awareness of stimulus differences, the *Awareness group* was told some stimuli may change slightly and participants needed to decide via mouse-click whether the stimulus matched one of the three stimuli that were known so far (i.e., 'identical' or 'different'). Moreover, participants were told they will be asked about stimulus differences in the end of the experiment. The *Control group* received neutral instructions (see SI, S3.1 for precise formulation) and indicated if the stimulus was a face or a geometrical shape (i.e., 'Face' or 'Form'). The rating appeared immediately after each trial with counterbalanced presentation side of both options. No time limit was set for the rating, which occurred after the 8 s stimulus duration for SCR scoring and US expectancy rating. However, participants were previously instructed to respond quickly and accurately. After finishing the second acquisition block, *extinction training* started without further instructions. Here, none of the stimuli was followed by an US anymore. Notably, in both groups, the CS+ A was replaced by six GSs A (analog with the Multiple GSs group in Experiment 1). The CS+ B and the CS- were, however, presented in their original version. After two blocks of extinction training (i.e., 12 extinction trials per stimulus), an immediate *test phase* was carried out. Here, in contrast to Experiment 1, all three 'original' stimuli were shown four times each in a randomized order.

Afterward, participants repeated valence and arousal ratings and were asked to indicate how many different faces and forms they have seen in total (open format). Aiming to increase reliability, written examples were provided, and participants could correct their ratings afterward if instructions had been misinterpreted (i.e., counting of trials instead of faces). Then, a second test phase with yet unknown stimuli (i.e., two novel face stimuli and one novel form; each shown two times) was carried out. Next, a task-specific memory test was conducted with all experimental stimuli (i.e., CS+ A; CS+ B; CS-; 12 GSs) being presented in randomized order. Participants rated their certainty, that each stimulus had been presented during the experimental task (continuous rating from −5 to +5). Of note, only six of the 12 GSs had been presented per participant due to counterbalancing of stimuli. Deviation from correct answers (i.e., +5 for stimuli that had been included; −5 for stimuli that had not been included) was computed for each participant and subtracted from the maximum value possible. The obtained Memory Accuracy Score has a range from 0 to 150, with a higher value indicating a more accurate identification of stimuli. In the end, ratings for the aversive US were assessed (e.g., valence, arousal, avoidance). After detachment of electrodes and completion of post-experimental questionnaires, all participants underwent a short version of a standardized Face Matching Test (Oxford Face Matching Test, Stantić et al., 2021; 2024). Here, in 200 trials (plus 12 attention check trials), two photographs were shown next to each other for 1600 ms. Afterward, participants decided if both pictures depicted the same or different *persons*. As in the first experiment, trial-wise US expectancy ratings were collected, and skin conductance activity was assessed continuously, during the 90 trials of the fear conditioning procedure. The processing of skin conductance data followed the first experiment. Again, detailed analyses are included in the SI, S3.2.

**Table 4**
Experimental design of experiment 2.

| Habituation Phase | Fear Acquisition Training | | Extinction Training | | Test Phase |
|---|---|---|---|---|---|
| | Block 1 | Block 2 | Block 1 | Block 2 | |
| CS+ A (2) | CS+ $A_{80\%}$ (5) | CS+ $A_{80\%}$ (5)* | [GS A 1–6] (1)* | [GS A 1–6] (1)* | CS+ A (4)* |
| CS+ B (2) | CS+ $B_{80\%}$ (5) | CS+ $B_{80\%}$ (5)* | CS+ B (6)* | CS+ B (6)* | CS+ B (4)* |
| CS- (2) | CS- (5) | CS- (5)* | CS- (6)* | CS- (6)* | CS- (4)* |

*Note.* CS+ A and CS+ B: neutral face stimuli (counterbalanced between groups); during Fear Acquisition, CS+ A and CS+ B were each reinforced by the aversive US (scream and matching fearful face) in 80 % of trials; CS+ A and CS+ B were not reinforced in any other phases; CS-: triangle form; * = trial-wise forced choice questions (Awareness: Identical or different one of the three original pictures? vs. Control: Form or Face?); A second test phase with novel stimuli (not included here) was applied after additional ratings in the end of the experiment. All face stimuli are part of the KDEF Set (Lundqvist et al., 2015). The following stimuli were used: AF19NES and AF07NES (CS+ s), AF19AFS and AF07AFS (USs); for the preparation of morph stimuli, the following stimuli were each individually morphed with each CS+: AF01NES, AF05NES, AF09NES, AF22NES, AF24NES, AF34NES.

*3.2.2. Data analyses*

*3.2.2.1. Fear learning, extinction training and test phases.* Data and analyses code are available via https://osf.io/qk2r6/. Separate LMMs were computed for each dependent measure (i.e., US expectancy ratings, SCR) and experimental phase. Here, *Stimulus* (CS+ A/GS A vs. CS+ B vs. CS-), *Trial Number* (as continuous linear predictor) and *Experimental Group* (Awareness vs. Control) were added as fixed factors. Interactions between fixed factors and random slopes per participant throughout Trials were included. Since US expectancy test phase model revealed a singular fit, only a random intercept was included. Additional robust analyses were performed if violation of model assumptions was noticed. The results overall align with the main models and can be reproduced via https://osf.io/qk2r6/.

Given the interest in the transitions between fear acquisition and extinction training (i.e., generalization decrement), and between extinction training and test phase (i.e., 'return' of US expectancy), cross-phase analyses with the fixed factors *Group*, *Stimulus* (CS+ A/GSs A vs. CS+ B vs. CS-) and *Trial Number* (as continuous predictor), and a random intercept per *Participant,* were calculated for both transitions separately. To specifically compare return of fear from the last extinction to the first test trial, preregistered slopes analyses for each Stimulus and Group were calculated.

*3.2.2.2. Analysis of accompanying ratings.* Accompanying ratings (i.e., number of faces; memory accuracy; trial-wise discrimination ratings; OFMT score) were compared between groups using independent samples *t-tests*. As preregistered, additional subgroup analyses were conducted: The Awareness group was split up regarding their ratings (i.e., performance in trial-wise discrimination; number of faces), and LMMs for extinction training and transition to test were repeated. To gain insights into affective stimulus ratings, valence and arousal ratings for the three conditioned stimuli at three timepoints were analyzed (see SI, S3.4).

*3.3. Results*

*3.3.1. Habituation and fear acquisition*

US expectancy data are depicted in Fig. 2A. During habituation trials, a significant main effect confirmed an overall decline for US expectancy ratings, $F(1, 78.01) = 26.49$, $p < .011$. Additionally, a main effect of Stimulus, $F(2, 312.01) = 9.26$, $p < .001$, indicated that both designated CS+ s (face stimuli) elicited higher ratings than the CS- (geometrical form), $t(312) = -3.93$, $p_{adj} < .001$ and $t(312) = -3.49$, $p_{adj} = .001$.

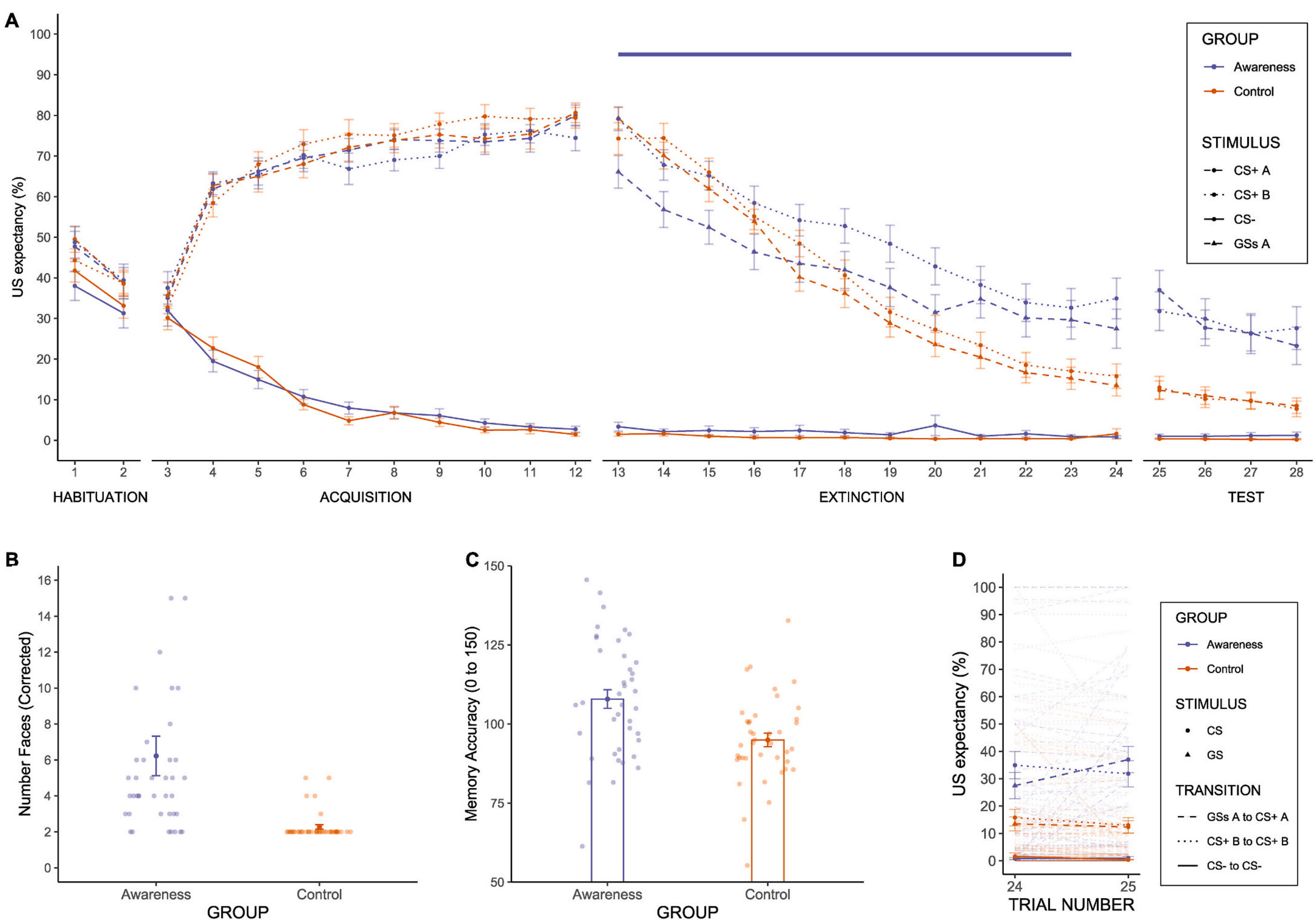


**Fig. 2.** Trial-wise mean US expectancy ratings (A) in the second experiment per group and stimuli. The violet line at the top of the extinction training phase indicates the trials for the Awareness group, in which US expectancy for the GSs A was significantly lower than for the CS+ B. Follow-up comparisons for the Control group did not show significant differences between GSs A and CS+ B. In (B) and (C), accompanying rating data of the second experiment are shown. Mean and standard deviation of (B) the number of different faces (open question, corrected ratings), and (B) the task-specific memory (Memory Accuracy Score, 0 to 150; a higher value indicates more accurate identification of known vs. unknown stimuli after the conditioning procedure), per group. Single, lighter dots represent individual scores/ ratings. In (D), the transition for US expectancy from the last extinction trial to the first test trial is depicted per stimulus type and group. Single participants reactions are depicted in the background. To enhance visibility, single plots per transition (i.e., GSs A to CS+ A, CS+ B to CS+ B, CS- to CS-) are provided in the SI, S3.7. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

These results might reflect an effect of the pre-experimental experience of a scream (later US), which is more naturally associated with face stimuli. Importantly, US expectancy for both later CS+ stimuli did not differ before start of fear acquisition, $t(312) = .43$, $p_{adj} = .664$. No effect involving the factor Group, nor any interaction effect showed significance ($ps \geq .366$, full statistics see SI, S3.2).

During fear acquisition training, both CS+s were paired with an aversive US. For US expectancy, a significant interaction effect of Stimulus and Trial Number was observed, $F(2, 2232) = 301.36$, $p < .001$. Follow-up tests confirmed a US expectancy increase for both CS+s (CS+ A: slope estimate = 9.63, $CI_{95}$[7.99; 11.27], $t(202) = 14.13$, $p_{adj} < .001$; CS+ B: slope estimate = 9.69, $CI_{95}$[8.04; 11.33], $t(202) = 14.21$, $p_{adj} < .001$), but a decrease for the CS- (slope estimate = −8.12, $CI_{95}$[−9.76; 11.27], $t(202)$ = -11.90, $p_{adj} < .001$; i.e., differential fear conditioning, see Fig. 2A). Importantly, there was no evidence for significant differences between both CS+s in any trials ($ps_{adj} = 1.000$). No significant main or interaction effects including the factor Group were revealed ($ps \leq .079$), suggesting comparable, successful differential fear learning at US expectancy level for both groups. Supporting US expectancy data, valence and arousal ratings for CSs (before habituation, after the first block of acquisition, after immediate test) suggested that CS+s were evaluated as more unpleasant and arousing than the CS- after pairing with the aversive US in both groups (Fig. S6). Detailed analyses are included in the SI, S3.4.

SCR data unexpectedly revealed a differential responding between groups towards the three stimuli during the initial habituation phase. However, there was no evidence that both groups showed differential fear learning at a physiological level. Given that differential acquisition of SCR to the CSs was not established, SCR data was not reported in the main text (the detailed results are included in the SI, S3.3).

#### 3.3.2. *Extinction training and test phase*

During extinction training, the CS+ A was replaced with multiple, highly similar GSs A, while the CS+ B and the CS- were presented in their original form. To investigate generalization decrement for GSs A, the transition between end of acquisition and beginning of extinction training was analyzed. Here, a significant three-way interaction between Group, Trial Number and Stimulus (CS+ A/GSs A vs. CS+ B vs. CS-) was apparent, $F(2, 390) = 5.79$, $p = .003$. Follow-up analysis revealed a significant decrease of US expectancy for the transition from CS+ A to the first GS A only for the Awareness group (slope estimate = −6.95, $CI_{95}$[−11.33; −2.56], $t(390) = -4.20$, $p_{adj} < .001$). The decrease for the Control group did not reach significance level (slope estimate = −.67, $CI_{95}$[−3.92; 2.58], $t(390) = -.41$, $p_{adj} = 1.000$). All other slope estimates in both groups appeared to be similarly flat, $ps_{adj} \geq .606$. Hence, results point towards a decrement of US expectancy for the GSs A only in the Awareness group.

The LMM for US expectancy across extinction training revealed a significant three-way interaction between Group, Trial Number and Stimulus, $F(2, 2712) = 27.56$, $p < .001$. Aligning with the absence of the US, US expectancy ratings significantly decreased for both CS+s in both groups ($ps_{adj} < .001$, see below), while ratings for the CS- remained stable at a very low level (slope $\text{estimate}_{\text{Awareness}} = -.57$, $CI_{95}$[−3.73; 2.59], $t(137) = -.49$, $p_{adj} = 1.000$, slope $\text{estimate}_{\text{Control}} = -.20$, $CI_{95}$[−3.36; 2.95], $t(137) = -.17$, $p_{adj} = 1.000$). The decrease for both the GSs A and the CS+ B was less steep for the Awareness group (slope $\text{estimate}_{\text{GSs A}} = -11.16$, $CI_{95}$[−14.31; −8.00]; slope $\text{estimate}_{\text{CS+ B}} = -14.02$, $CI_{95}$[−17.18;-10.86), compared to the Control group (slope $\text{estimate}_{\text{GSs A}} = -21.23$, $CI_{95}$[−24.38; −18.07]; slope $\text{estimate}_{\text{CS+ B}} = -20.80$, $CI_{95}$[−23.96; −17.64], $t_{\text{GSs A}}(137) = 6.04$, $p_{adj} < .001$ and $t_{\text{CS+ B}}(137) = 4.06$, $p_{adj} < .001$). Groupwise follow-up tests comparing EMMs for US expectancy ratings for the GSs A and the CS+ B in each trial were applied. Significant differences between GSs A and CS+ B are depicted at the top of Fig. 2A (i.e., horizontal line): For the Awareness group, US expectancy ratings for the GSs A were smaller during *all* but the last extinction trials ($ps_{adj} \leq .016$). For the Control group, however, there was no evidence for significant different US expectancy ratings of GSs A and CS+ B ($ps_{adj} \geq .098$, see Table S8). In sum, only the Awareness group showed higher US expectancy for the CS+ B compared to the GSs A, presumably due to a generalization decrement caused by awareness of differences between stimuli. At the same time, decline of US expectancy was less steep for both the GSs A and the CS+ B in comparison with the Control group.

Immediately after the end of extinction training, the three conditioned stimuli were presented four times in their original form (*test phase*). The transition between the end of extinction training and the first test trial was of specific interest. Here, a significant interaction between Stimulus (GSs A/CS+ A vs. CS+ B vs. CS-) and Group, $F(2, 390) = 32.71$, $p < .001$, was observed. Averaged across the two phases, the Awareness group showed significantly higher overall US expectancy for GSs A/CS+ A, $t(127) = 4.88$, $p_{adj} < .001$, and CS+ B, $t(127) = 4.80$, $p_{adj} < .001$, compared with the Control group. However, there was no group difference for the CS-, $t(127) = -.02$, $p_{adj} = .984$. The three-way interaction between Trial, Group and Stimulus did not reach significance though, $F(2, 390) = 1.68$, $p = .188$. Furthermore, none of the other main effects or interactions revealed significant results (all $ps \geq .080$). As preregistered, slope estimates for each stimulus type and group were tested. As depicted in Fig. 2D, a significant increase for the transition between the last GS A in extinction training and the first CS+ A test trial was observed only in the Awareness group (slope estimate = 4.76, $CI_{95}$[1.59; 7.94], $t(390) = 2.95$, $p_{adj} = .020$). This pattern was, however, not revealed in the Control group (slope estimate = −.58, $CI_{95}$[−3.75; 2.60], $t(390) = -.36$, $p_{adj} = 1.000$). Similarly, US expectancy ratings for the CS+ B and the CS- did not significantly change in none of the groups across the two phases (CS+ B: slope $\text{estimate}_{\text{Awareness}} = -1.57$, $CI_{95}$[−5.20; 2.07], slope $\text{estimate}_{\text{Control}} = -.20$, $CI_{95}$[−2.88; 2.47]; CS-: slope $\text{estimate}_{\text{Awareness}} = .08$, $CI_{95}$[−3.10; 3.25], slope $\text{estimate}_{\text{Control}} = -.61$, $CI_{95}$[−3.79; 2.56]; all $ps_{adj} = 1.000$).

During the test phase, the three-way interaction between Group, Trial Number, and Stimulus did not reach significance, $F(2, 870) = 1.82$, $p = .162$. However, there was a significant interaction between Group and Stimulus, $F(2, 870) = 49.27$, $p < .001$. Follow-up comparisons revealed significantly higher US expectancy ratings for both the CS+ A, $t(99.8) = 5.24$, $p_{adj} < .001$, and the CS+ B, $t(99.8) = 5.39$, $p_{adj} < .001$, in the Awareness compared to Control group. In contrast, CS- ratings did not differ between groups, $t(99.8) = .23$, $p_{adj} < .816$. Overall higher US expectancy ratings for in the Awareness group during test phase directly connect with results from the extinction training, where this group showed flatter extinction learning gradients for both GSs A and the CS+ B. Additionally, a significant interaction between Stimulus and Trial Number was observed, $F(2, 870) = 4.62$, $p = .010$. Follow-up analyses showed that the CS+s but not the CS- showed decreasing US expectancy throughout test trials (slope $\text{estimate}_{\text{CS+ A}} = -3.10$, $CI_{95}$[−4.52; −1.67], $t(870) = -4.26$, $p_{adj} < .001$; slope $\text{estimate}_{\text{CS+ B}} = -1.81$, $CI_{95}$[−3.24; −.38], $t(870) = -2.49$, $p_{adj} = .026$; slope $\text{estimate}_{\text{CS-}} = -1.81$, $CI_{95}$[−3.24; −.38], $t(870) = -.02$, $p_{adj} = .987$). This decrease can be interpreted as further extinction learning in absence of aversive events. Exploratory analyses for novel stimuli (NS) in a second test phase may tentatively suggest no difference in the initial reaction towards novel face stimuli (i.e., original face stimuli from the same dataset) between groups, but a steeper decrease of US expectancy from the first to the second NS test trial in the Awareness compared with the Control group, $t(78) = -2.77$, $p = .007$. Detailed results and interpretation are included in the SI, S3.6.

#### 3.3.3. *Accompanying rating data – manipulation check and subgroup analyses*

To evaluate the between group manipulation, additional rating data was analyzed. At first, after end of the test phase, participants indicated how many different faces and forms they had seen (open format). The Awareness group indicated a mean of 7.00 (SD = 12.2) different faces, while the Control group reported 4.63 (SD = 12.1), $t(78) = .85$, $p = .400$.

Taking corrected answers (e.g., if trials instead of faces/forms had been counted) into account, the Awareness group reported 6.23 (SD = 6.97) and the Control group 2.28 (SD = .78) faces, $t(78) = 3.60$, $p = .001$ (see Fig. 2B). The reported amount of geometrical forms did not differ between groups in uncorrected ratings, $t(78) = -1.15$, $p = .252$, and was completely accurate (i.e. one form) in both groups for corrected ratings. Secondly, post-experimental memory ratings for each stimulus (i.e., certainty, if this stimulus had been included in the experiment) were more accurate in the Awareness (M = 107.9, SD = 18.3) compared to the Control group (M = 95.0, SD = 13.4), $t(78) = -3.56$, $p = .001$ (Fig. 2C). In sum, rating data can be interpreted as a manipulation check and point towards a better detection of stimulus differences in the Awareness group through experimental manipulation.

Furthermore, participants answered trial-wise forced choice questions *during* the fear conditioning procedure (i.e., identical vs. different stimulus; form vs. face). Here, especially the performance of the Awareness group during extinction training was of interest to investigate individual differences in stimulus discrimination. While discrimination accuracy in the Awareness group across all phases was high (75.50 %; M = 30.20, SD = 3.33), participants correctly recognized GSs A extinction trials as being different from the CS+ A in 66.8 % (M = 8.02, SD = 3.37, see Fig. 3A). In the Control group, where participants decided if a face or form was presented, overall discrimination accuracy was very high (89.75 % correct answers; M = 35.90, SD = .26).

For preregistered subgroup analyses, the Awareness group was split based on trial-wise discrimination accuracy for GSs A *during* extinction training. The preregistered cut-off (i.e., 50 % accuracy) was too low to result in reasonable subgroups (i.e., 10 vs. 30 participants). Therefore, subgroups were defined via median split (median = 8.5; $M_{\text{HigherAccuracy}} = 11.00$, $SD = 1.07$; $M_{\text{LowerAccuracy}} = 5.10$, $SD = 2.10$). A significant three-way interaction between the factors Subgroup, Trial Number and Stimulus pointed towards an stronger differentiation of US expectancy between the GSs A and the CS+ B during extinction training in participants with higher discrimination accuracy, $F(2, 1352) = 11.50$, $p < .001$ (see Fig. 3B). Follow-up tests suggest that the subgroup with higher discrimination accuracy indicated less US expectancy for the GSs A in comparison to CS+ B in all extinction trials (all $ps_{\text{adj}} \leq .001$), while the subgroup with lower accuracy significantly differed in the earlier (all $ps_{\text{adj}} \leq .037$) but not the later trials (all $ps_{\text{adj}} \geq .370$, see horizontal lines in Fig. 3B). Surprisingly, the lower accuracy subgroup showed overall flatter extinction learning slopes both for GSs A, $t(66.3) = -3.72$, $p_{\text{adj}} = .007$, and the CS+ B, $t(66.3) = -2.45$, $p_{\text{adj}} = .034$, compared with the higher accuracy subgroup. Slope analyses for the transition from extinction training to test (i.e., from GSs A to CS+ A) showed a significant increase for the higher accuracy (slope estimate = 8.12, $CI_{95}[2.64; 13.59]$, $t(190) = 2.93$, $p_{\text{adj}} = .023$), but not the lower accuracy group (slope estimate = 1.41, $CI_{95}[-4.06; 6.88]$, $t(190) = .51$, $p_{\text{adj}} = 1.000$). However, as in the main analyses, the three-way interaction between Subgroup, Stimulus and Trial Number missed significance here, $F(2, 190) = 1.14$, $p = .321$. Notably, no differences between subgroups for the CS- had been expected or were observed, whereas the inclusion of the CS- in analyses may increase the risk of Type II error. Therefore, an exploratory model for the transition between extinction training and test, excluding CS- trials, was run. However, this model yielded similar, non-significant results for the three-way interaction, $F(2, 114) = 2.90$, $p = .091$. Detailed analyses are included in the SI, S3.5.1. Overall, the subgroup results tentatively indicate that a more accurate detection of *GSs* may be associated with stronger differentiation of US expectancy between extinction training stimuli, and a stronger increase of US expectancy in test (i.e., less GS extinction generalization). Additionally, also the retrospectively indicated number of different faces (i.e., open question) was used for subgroup analyses (preregistered cut-off = 5 faces). Accordingly, a higher number of indicated faces tends to be associated with overall lower US expectancy for GSs A during extinction training, and a descriptively stronger return of US expectancy when confronted with the original CS+ A (see SI, S3.5).

### 3.4. Discussion of experiment 2

Following the results of the first experiment, we investigated the influence of manipulated awareness of physical stimulus differences on extinction learning and its generalization. To increase awareness of differences, the Awareness group received corresponding instructions and needed to decide trial-wise, if the stimulus was identical to one of the original stimuli. As expected, only the Awareness group showed a decrease in US expectancy for the GSs (i.e., generalization decrement) in the beginning of extinction training. Consistently, throughout extinction training, US expectancy ratings for the CS+ and the GSs differed more strongly (i.e., lower ratings for the GSs) in the Awareness group compared with the Control group. Moreover, the Awareness group

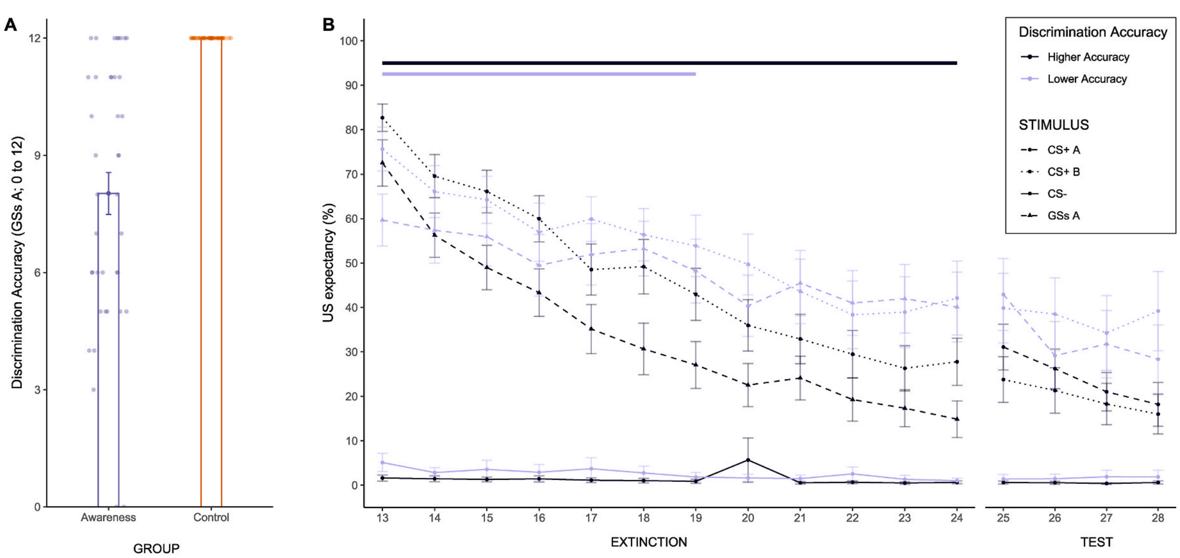


**Fig. 3.** (A) Mean and standard deviation of the performance during the trial-wise forced-choice task in the extinction trials showing GSs A, for Awareness and Control group. Single, lighter dots represent individual scores/ratings. In (B), US expectancy ratings for extinction training and test phase for median split subgroups of the Awareness group (higher vs. lower accuracy in discrimination of GSs A in trial-wise ratings) are shown. Horizontal lines at the top of the graph indicate trial-wise significant US expectancy differences between GSs A and CS+ B for each subgroup.

showed generally flatter extinction learning gradients for both the CS+ B and GSs A, and, accordingly, higher remaining US expectancy in the following test phase trials. Importantly, as hypothesized, an increase of US expectancy after the end of extinction training for the formerly replaced CS+ A (i.e., change from GSs A to CS+ A) did *only* appear in the Awareness but not in the Control group. This pattern is therefore in line with findings about limited generalization of GS extinction learning back to the original feared stimulus, resulting in a return of fear (e.g., Vervliet et al., 2004, 2005; Vervoort et al., 2014; Wong & Lovibond, 2020).

As presumed, our results suggest that the awareness for stimulus differences was successfully influenced "top-down". Here, not only group differences during extinction learning and generalization, but also task-specific ratings (i.e., number of different faces, memory accuracy) point towards a successful awareness induction. Potentially, the experimental manipulation resulted in higher attention allocation during stimulus observation, leading to a better detection and recall of visual differences (e.g., Kim & Chong, 2022). Hence, an attention shift towards differences between GSs A and CS+ A (i.e., non-shared, unique GSs features) may have triggered weaker extinction learning for features that are shared between GS and CS+ (see Barry et al., 2015, 2016, 2017; Vervliet et al., 2005). The later confrontation with the original CS+ A therefore resulted in an increase of US expectancy. As participants in the Control group were presumably less aware of stimulus differences, GS extinction learning was not impeded, and no return of US expectancy was visible. We could therefore show that, after increasing the attention for stimulus differences, also highly similar stimuli can be differentiated successfully and threat expectancies are altered accordingly. Aligning with between-group results, subgroup analyses tentatively suggest that a higher discrimination accuracy for GSs (i.e., trial-wise task) may be associated with stronger return of US expectancy. However, participants that showed lower accuracy in recognition of GSs exhibited flatter extinction gradients, which may result from overall lower attention in that subgroup. Furthermore, exploratory trial-by-trial analyses on the association between correct recognition of a stimulus ("same" vs. "different") and US expectancy showed that the misinterpretation of GS as the CS+ was associated with higher US expectancy within the same trial, which supports previous findings (Struyf et al., 2017; Zaman, Struyf, et al., 2019; for details, see SI, S3.5.2).

Beyond the influence on GS extinction learning, surprisingly, the Awareness group showed slower extinction learning (i.e., flatter gradient) also for the not-replaced CS+ B. Hence, the awareness manipulation seems to have influenced *both* GS and CS+ extinction learning. Generally, awareness instructions may lead to an overall assumption that each original stimulus will potentially change over the course of the task. While the consequentially increased attention for differences may represents a precondition for a (mainly) successful differentiation between GSs A and CS+ A, it could have a disconcerting effect for the evaluation of the CS+ B, which was never replaced. The sequential presentation of stimuli with long lasting ITI may additionally fostered uncertainty or biases during evaluation (e.g., the CS+ B was misclassified as being a GS in 13.33 % of extinction trials). At the same time, group differences in awareness may have led to divergent generalization effects during extinction learning: In the Control group, where the GSs A were presumably perceived as the original CS+ A, learning that the (assumed) CS+ A was no longer followed by the aversive US may have generalized also towards the CS+ B (and vice versa), potentially facilitating faster extinction learning. In the Awareness group, where GSs A were better recognized as being different from CS+ A, extinction learning for GSs A may not have generalized as strongly to CS+ B. This reduced generalization could have contributed to impeded extinction learning for CS+ B in the Awareness group compared to the Control group. Notably, however, differences in task difficulty or cognitive load during extinction training may cause slower extinction learning as well (e.g., Cheng et al., 2022; Kamphuis & Telch, 2000; Raes et al., 2009). In our study, trial-wise rating results suggest that the task was more difficult for the Awareness group than the Control group. However, the better differentiation between GSs and CS+ during extinction training cannot be sufficiently explained by task difficulty effects. Hence, application of a comparably demanding control task may help to disentangle (potentially interacting) effects of attention allocation and cognitive load.

Of note, the generalization decrement at the start of extinction training (i.e., from CS+ A to GS A) showed a more pronounced between-group effect compared to the return of US expectancy after extinction training (i.e., from GS A to CS+ A). One explanation may be the use of multiple GSs replacing the CS+ A, a strategy suggested to improve GS extinction generalization (e.g., Rowe & Craske, 1998; Shiban et al., 2015). Using a *single* GS might have resulted in a stronger increase of US expectancy in the Awareness group. Furthermore, exploratory findings may tentatively indicate faster extinction learning (i.e., steeper decrease of US expectancy) for novel face stimuli in the Awareness compared to the Control group (see SI, S3.6). This might be interpreted as differential effects of awareness on generalization: While awareness for stimulus differences seems to impede GS extinction generalization towards the original CS+ (i.e., increase of US expectancy), it may facilitate extinction learning for yet unknown, less similar stimuli from the same category (i.e. female faces). Here, the use of *multiple, different GSs* during extinction training may have supported the transfer of learning in situations with other generally related stimuli. Nonetheless, given that the test phase with novel stimuli followed a potentially interfering rating task, and that the similarity of stimuli had not been controlled for, these exploratory results should be interpreted with caution. Future studies may specifically investigate the effect of awareness in generalization to novel stimuli.

## 4. General discussion

In two experiments, we first aimed to investigate the effect of extinction learning with multiple, highly similar GS(s) as a potential strategy for improving GS extinction generalization. However, we found no group differences in GS extinction generalization, presumably because participants were not aware of the perceptual differences between the stimuli presented during extinction training. In a second experiment, we therefore aimed to further explore the effect of awareness of stimulus differences on (GS) extinction learning and generalization. Results indicate that the CS+ and highly similar GSs *can* be distinguished after applying additional instructions and tasks. Accordingly, attention allocation towards stimulus differences significantly altered (GS) extinction learning and resulted in a return of US expectancy when confronted with the original CS+. Thus, paradoxically, a more accurate stimulus differentiation is associated with worse generalization of GS extinction learning, reflected by a strong increase of threat expectancy.

Combining both studies, the findings of our second experiment presumably support a pivotal role of missing awareness to explain the lack of group differences in the first experiment. Beyond that, differences between initial validation and the two experiments highlight that physically similar stimuli may be differentiated in a context not involving any threat or aversive stimuli (i.e., the validation task), but, under application of additional instructions and ratings, also in a fear conditioning situation (i.e., second experiment). Therefore, mechanisms for the processing and evaluation of stimuli, for example attention allocation and cognitive load, may be altered by global contextual factors (e.g., experimental task, perceived threat) and in turn influence GS extinction learning beyond a sole effect of physical stimulus similarity. Especially in Experiment 2, a-priori assumptions and cognitive rules drawn from instructions and tasks may have influenced these processes, as they were explicitly manipulated. Several studies showed that individual beliefs about the underlying structure of a threat can also influence (GS) extinction learning and generalization (Lovibond et al., 2020; Struyf et al., 2018; Wong et al., 2020). An explicit and standardized assessment of threat beliefs throughout future experiments may gain additional insights into the role of individual as well as top-down influenced cognitive processes.

Our results emphasize the importance of accompanying processes during exposure: Presumably, therapeutic conditions as instructions and knowledge about therapy mechanisms, attention allocation and higher-order cognitive processes (i.e., meta-beliefs) can pivotally influence experiences and learning also during exposure therapy. Therefore, a careful and joint consideration and integration of these variables during treatment at an individual level may foster its effectiveness. For example, especially emphasizing joint features that one exposure exercise shares with other, also threat related situations, may be beneficial. Furthermore, explicitly addressing accompanying processes and factors may also increase the validity of experimental research, helping to further optimize treatment.

Several limitations of the present studies need to be discussed. At first, while the first experiment also investigated spontaneous recovery and reinstatement effects 24h later, we did not assess the longer-term stability of between-group effects in the second experiment. Testing the reactions toward the CSs at later timepoints would be of interest, especially as a long-term change of treat expectancy shall be achieved during treatment. Second, in contrast to US expectancy and affective ratings, SCR results in the Experiment 2 did not support differential fear acquisition at physiological level. Therefore, prerequisites for the interpretation of SCR data in the subsequent phases were not met. Descriptively, the Control group showed surprisingly strong SCR for the safety stimulus CS-, which potentially have been influenced by the trial-wise task for this group (i.e., face vs. form). Thus, a stronger orienting response towards the less frequent CS- may have impeded differential fear acquisition (e.g., Barry et al., 2023). However, also in Experiment 1, SCR results pointed towards unexpected group differences during fear acquisition. Compared with electric stimulation as a US, a loud scream was shown to result in faster SCR habituating and lower differential conditioning (Ney et al., 2023). The application of a different US may help to obtain more conclusive results at physiological level in future studies. Additionally, the assessment of US expectancy ratings during the presentation of the stimulus may have impeded the interpretability of SCR data. Third, although the results of the Control group in Experiment 2 (i.e., without awareness manipulation) are consistent with the findings of Experiment 1, direct comparisons should be done cautiously. For instance, the additional forced-choice task and stimuli (i.e., CS+ B and CS-) in Experiment 2 might have limited overall cognitive resources for stimulus discrimination. Notably, the absence of a safety cue in Experiment 1 was intended to avoid unintentional extinction effects due to overlapping characteristics between CS+ and CS- (see Chapter 2.2.3). However, interpretability of results was partly limited as reactions toward the CS+ could not be directly compared with a safety stimulus. Thus, Experiment 2 introduced a *categorically unrelated* CS-, aiming to minimize potential interference. However, we cannot be certain whether an introduction of a CS- in Experiment 1 would have altered the results as well. Furthermore, in Experiment 1, presentation of novel stimuli in the beginning of the immediate test phase resulted in a strong orienting response and may have been interpreted as signaling an overall change in the contingencies, potentially influencing subsequent reactions to the CS+. Therefore, no NS was included in the immediate test phase of the second experiment, though doing so reduced comparability. Future studies specifically designed to test fear reactions towards novel stimuli (e.g., to model real life situations) may aim to control for novelty of stimuli. Fourth, we decided for the use of same-sex face stimuli (i.e., all pictures showed female faces) aiming to be able to create consistent GSs for both CS+ stimuli and to reduce potential effects of the participants gender (e.g., own-gender bias for the encoding of face stimuli; see Herlitz & Lovén, 2013). However, future studies may investigate how results could be affected by using different facial stimuli. Fifth, given that higher-order cognitive processes (e.g., meta threat beliefs) about the structure of stimuli and threats can influence extinction, their structured assessment during and after the conditioning procedure (e.g., via interviews) could be insightful to further disentangle effects of attention allocation, cognitive processes and perception. Notably, exploratory investigation of the relationship between general face matching performance (OFMT, Stantić et al., 2024) and task-specific discrimination and retrospective memory accuracy did not reveal significant associations ($rs = -.02$ to $.21$, $ps \geq .188$), suggesting that overall face recognition ability in a neurotypical sample may not be crucial for stimulus differentiation in our task. Thus, assessment of specific strategies to differentiate between highly similar (face) stimuli, and a replication using different stimuli, would be of interest. The explicit assessment of stimulus differentiation *within* the fear conditioning procedure, as in our second study, can allow to examine the influence of failed perceptual discrimination on extinction learning (e.g., Struyf et al., 2017; Zaman, Ceulemans, et al., 2019; Zaman, Struyf, et al., 2019). However, the limited sample size (i.e., Awareness group only) does only allow preliminary interpretations of subgroup data. Though, replicating the findings in a larger sample seems necessary.

In conclusion, our first experiment did not show differences in extinction learning and its generalization despite the use of physically different extinction stimuli. Thus, surprisingly, no limited generalization of single GS extinction learning nor its improvement through increased stimulus variability (i.e., multiple GSs) was observed. Missing discrimination of the highly similar stimuli needs to be considered as a reasonable explanation for the results. The results of the second experiment, however, suggest that additional instructions and forced-choice questions were able to presumably increase awareness of physical stimulus differences. This, in turn, resulted in successful differentiation between GSs and CS+ during extinction training, and a return of US expectancy for the originally feared stimulus after GSs extinction training with multiple GSs (i.e., limited GS extinction generalization). Furthermore, in the Awareness group, extinction learning for the GSs but also the other CS+ appeared to be slower, resulting in flatter gradients and higher remaining US expectancy in test. Our results point out that, although stimuli with the same level of perceptual similarity were used, awareness of stimulus differences can strongly influence extinction learning and generalization. As GS extinction generalization may play a pivotal role in the long-term learning from exposure therapy in anxiety disorders, future studies investigating the influence of awareness of perceptual differences in interaction with attentional and higher-order cognitive processes may further inform and optimize treatment.

## CRediT authorship contribution statement

**Paula Engelke:** Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Alex H.K. Wong:** Writing – review & editing, Methodology, Conceptualization. **Andre Pittig:** Writing – review & editing, Supervision, Methodology, Conceptualization.

## Funding

Paula Engelke was supported by a scholarship by the Studienstiftung des Deutschen Volkes (German National Academic Foundation). Data collection was supported by overhead funding and gender equality support funding of the Faculty of Human Sciences, University of Würzburg, and by gender equality support funding of the Institute of Psychology, University of Würzburg, to Paula Engelke.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Acknowledgements

The authors would like to thank Pia Burghartz, Judith Dehnen, Elena Duscher, Sevval Eryüksel, Maria Kirchner, Lara Scheuer and Isabel Stangier for their help in data collection and processing. Additionally, we sincerely thank Mirta Stantić for providing materials and valuable advice on implementing the Oxford Face Matching Test.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.brat.2025.104787.

## Data availability

Data and analytic code for this study are available via https://osf.io/qk2r6/.

## References

Barry, T. J., Griffith, J., Vervliet, B., & Hermans, D. (2016). The role of stimulus specificity and attention in the generalization of extinction. *Journal of Experimental Psychopathology, 7*, 143–152. https://doi.org/10.5127/jep.048615

Barry, R. J., Steiner-Lim, G. Z., Cave, A. E., De Blasio, F. M., & MacDonald, B. (2023). Effects of interstimulus interval and significance on electrodermal and central measures of the phasic orienting reflex (OR) in a dishabituation task. *Scientific Reports, 13*(1), Article 13546. https://doi.org/10.1038/s41598-023-40428-7

Barry, T. J., Vervliet, B., & Hermans, D. (2015). An integrative review of attention biases and their contribution to treatment for anxiety disorders. *Frontiers in Psychology, 6* (July). https://doi.org/10.3389/fpsyg.2015.00968. Article July.

Barry, T. J., Vervliet, B., & Hermans, D. (2017). Feature specific attention and return of fear after extinction. *Journal of Experimental Psychopathology, 8*, 76–87. https://doi.org/10.5127/jep.051115

Benito, K., Pittig, A., Abramowitz, J., Arch, J. J., Chavira, D., de Kleine, R., De Nadai, A. S., Hermans, D., Hofmann, S. G., Hoyer, J., Huppert, J. D., Kircanski, K., McEvoy, P. M., Meyer, H., Monfils, M.-H., Papini, S., Rief, W., Rosenfield, D., Storch, E. A., … Smits, J. A. J. (2024). Mechanisms of change in exposure therapy for anxiety and related disorders: A research agenda. *Clinical Psychological Science*. , Article 21677026241240730. https://doi.org/10.1177/21677026241240727

Boschen, M. J., Neumann, D. L., Waters, A. M., Boschen, M. J., Neumann, D. L., & Waters, A. M. (2009). Relapse of successfully treated anxiety and fear: Theoretical issues and recommendations for clinical practice. *Australian and New Zealand Journal of Psychiatry, 43*(2), 89–100. https://doi.org/10.1080/00048670802607154

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of pavlovian learning. *Psychological Bulletin, 114*(1), 80–99. https://doi.org/10.1037//0033-2909.114.1.80

Bouton, M. E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry, 52*(10), 976–986. https://doi.org/10.1016/S0006-3223(02)01546-9

Carpenter, J. K., Andrews, L. A., Witcraft, S. M., Powers, M. B., Smits, J. A. J., & Hofmann, S. G. (2018). Cognitive behavioral therapy for anxiety and related disorders: A meta-analysis of randomized placebo-controlled trials. *Depression and Anxiety, 35*(6), 502–514. https://doi.org/10.1002/da.22728

Cheng, Y., Jackson, T. B., & MacNamara, A. (2022). Modulation of threat extinction by working memory load: An event-related potential study. *Behaviour Research and Therapy, 150*, Article 104031. https://doi.org/10.1016/j.brat.2022.104031

Connor, K. M., Davidson, J. R. T., Churchill, L. E., Sherwood, A., Weisler, R. H., & Foa, E. (2000). Psychometric properties of the social phobia inventory (SPIN): New self-rating scale. *British Journal of Psychiatry, 176*(4), 379–386. https://doi.org/10.1192/bjp.176.4.379

Costa, P. T., & McCrae, R. R. (1992). The five-factor model of personality and its relevance to personality disorders. *Journal of Personality Disorders, 6*(4), 343–359. https://doi.org/10.1521/pedi.1992.6.4.343

Craske, M. G., Hermans, D., & Vervliet, B. (2018). State-of-the-art and future directions for extinction as a translational model for fear and anxiety. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1742), Article 20170025. https://doi.org/10.1098/rstb.2017.0025

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy, 58*, 10–23. https://doi.org/10.1016/j.brat.2014.04.006

Craske, M. G., Treanor, M., Zbozinek, T. D., & Vervliet, B. (2022). Optimizing exposure therapy with an inhibitory retrieval approach and the OptEx Nexus. *Behaviour Research and Therapy, 152*, Article 104069. https://doi.org/10.1016/j.brat.2022.104069

Gerdes, A. B. M., Fraunfelter, L., & Alpers, G. W. (2020). Hear it, fear it: Fear generalizes from conditioned pictures to semantically related sounds. *Journal of Anxiety Disorders, 69*. https://doi.org/10.1016/j.janxdis.2019.102174. November 2019), Article November 2019.

Gerlach, A. L., Andor, T., & Patzelt, J. (2008). Die Bedeutung von Unsicherheits-Intoleranz für die generalisierte Angststörung: Modellüberlegungen und Entwicklung einer deutschen Version der Unsicherheitsintoleranz-Skala. *Zeitschrift für Klinische Psychologie und Psychotherapie, 37*(3), 190–199. https://doi.org/10.1026/1616-3443.37.3.190

Haaker, J., Golkar, A., Hermans, D., & Lonsdorf, T. B. (2014). A review on human reinstatement studies: An overview and methodological challenges. *Learning & Memory, 21*(9), 424–440. https://doi.org/10.1101/lm.036053.114

Hartig, F. (2022). DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models Version v0.4.6. http://florianhartig.github.io/DHARMa/.

Hautzinger, M., Bailer, M., Hofmeister, D., & Keller, F. (2012). *Allgemeine Depressionsskala (ADS) (2., überarbeitete und neu normierte Auflage)*. Hogrefe.

Hayes, A. F., Glynn, C. J., & Huge, M. E. (2012). Cautions regarding the interpretation of regression coefficients and hypothesis tests in linear models with interactions. *Communication Methods and Measures, 6*(1), 1–11. https://doi.org/10.1080/19312458.2012.651415

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition, 21*(9–10), 1306–1336. https://doi.org/10.1080/13506285.2013.823140

Herzog, K., Andreatta, M., Schneider, K., Schiele, M. A., Domschke, K., Romanos, M., Deckert, J., & Pauli, P. (2021). Reducing generalization of conditioned fear: Beneficial impact of fear relevance and feedback in discrimination training. *Frontiers in Psychology, 12*, Article 665711. https://doi.org/10.3389/fpsyg.2021.665711

Hoyer, J., Čolić, J., Pittig, A., Crawcour, S., Moeser, M., Ginzburg, D., Lin, J., Wiltink, J., Leibing, E., & Stangier, U. (2017). Manualized cognitive therapy versus cognitive-behavioral treatment-as-usual for social anxiety disorder in routine practice: A cluster-randomized controlled trial. *Behaviour Research and Therapy, 95*, 87–98. https://doi.org/10.1016/j.brat.2017.05.012

Kamphuis, J. H., & Telch, M. J. (2000). Effects of distraction and guided threat reappraisal on fear reduction during exposure-based treatments for specific fears. *Behaviour Research and Therapy, 38*(12), 1163–1181. https://doi.org/10.1016/S0005-7967(99)00147-3

Kim, C., & Chong, S. C. (2022). Effects of attention on visual discrimination and awareness: Evidence from eye-specific attention. *Journal of Experimental Psychology: Human Perception and Performance, 48*(7), 743–754. https://doi.org/10.1037/xhp0001019

Kircanski, K., Mortazavi, A., Castriotta, N., Baker, A. S., Mystkowski, J. L., Yi, R., & Craske, M. G. (2012). Challenges to the traditional exposure paradigm: Variability in exposure therapy for contamination fears. *Journal of Behavior Therapy and Experimental Psychiatry, 43*(2), 745–751. https://doi.org/10.1016/j.jbtep.2011.10.010

Klein, Z., Ginat-Frolich, R., Barry, T. J., & Shechner, T. (2021). Effects of increased attention allocation to threat and safety stimuli on fear extinction and its recall. *Journal of Behavior Therapy and Experimental Psychiatry, 72*, Article 101640. https://doi.org/10.1016/j.jbtep.2021.101640

Koller, M. (2016). robustlm: An R package for robust Estimation of linear mixed-effects models. *Journal of Statistical Software, 75*(6). https://doi.org/10.18637/jss.v075.i06

Lang, A. J., & Craske, M. G. (2000). Manipulations of exposure-based therapy to reduce return of fear: A replication. *Behaviour Research and Therapy, 38*(1), 1–12. https://doi.org/10.1016/S0005-7967(99)00031-5

Lau, J. Y. F., Lissek, S., Nelson, E. E., Lee, Y., Roberson-Nay, R., Poeth, K., Jenness, J., Ernst, M., Grillon, C., & Pine, D. S. (2008). Fear conditioning in adolescents with anxiety disorders: Results from a novel experimental paradigm. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*(1), 94–102. https://doi.org/10.1097/chi.0b01e31815a5f01

Lenth, R. V. (2023). *emmeans: Estimated marginal means, aka least-squares means* [R] Version v1.8.9. .

Lipp, O. V., Waters, A. M., Luck, C. C., Ryan, K. M., & Michelle, G. (2020). Novel approaches for strengthening human fear extinction: The roles of novelty, additional USs, and additional GSs. *Behaviour Research and Therapy*. , Article 103529. https://doi.org/10.1016/j.brat.2019.103529

Lovibond, P. F., Lee, J. C., & Hayes, B. K. (2020). Stimulus discriminability and induction as independent components of generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(6), 1106–1120. https://doi.org/10.1037/xlm0000779

Lundqvist, D., Flykt, A., & Öhman, A. *Karolinska directed emotional faces*. (2015). American Psychological Association. https://doi.org/10.1037/t27732-000.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Mertens, G., Boddez, Y., Sevenster, D., Engelhard, I. M., & De Houwer, J. (2018). A review on the effects of verbal instructions in human fear conditioning: Empirical findings, theoretical considerations, and future directions. *Biological Psychology, 137*, 49–64. https://doi.org/10.1016/j.biopsycho.2018.07.002

Ney, L. J., Nichols, D. S., & Lipp, O. V. (2023). Fear conditioning depends on the nature of the unconditional stimulus and may be related to hair levels of endocannabinoids. *Psychophysiology, 60*(8), Article e14297. https://doi.org/10.1111/psyp.14297

Pittig, A., Heinig, I., Goerigk, S., Richter, J., Hollandt, M., Lueken, U., Pauli, P., Deckert, J., Kircher, T., Straube, B., Neudeck, P., Koelkebeck, K., Dannlowski, U., Arolt, V., Fydrich, T., Fehm, L., Ströhle, A., Totzeck, C., Margraf, J., … Wittchen, H.-U. (2023). Change of threat expectancy as mechanism of exposure-based psychotherapy for anxiety disorders: Evidence from 8,484 exposure exercises of 605 patients. *Clinical Psychological Science, 11*(2), 199–217. https://doi.org/10.1177/21677026221101379

Pittig, A., Heinig, I., Goerigk, S., Thiel, F., Hummel, K., Scholl, L., Deckert, J., Pauli, P., Domschke, K., Lueken, U., Fydrich, T., Fehm, L., Plag, J., Ströhle, A., Kircher, T., Straube, B., Rief, W., Koelkebeck, K., Arolt, V., … Wittchen, H. U. (2021). Efficacy of temporally intensified exposure for anxiety disorders: A multicenter randomized clinical trial. *Depression and Anxiety, 38*(11), 1169–1181. https://doi.org/10.1002/da.23204

Pittig, A., Treanor, M., LeBeau, R. T., & Craske, M. G. (2018). The role of associative fear and avoidance learning in anxiety disorders: Gaps and directions for future research. *Neuroscience & Biobehavioral Reviews, 88*, 117–140. https://doi.org/10.1016/j.neubiorev.2018.03.015

Pittig, A., Van Den Berg, L., & Vervliet, B. (2016). The key role of extinction learning in anxiety disorders: Behavioral strategies to enhance exposure-based treatments. *Current Opinion in Psychiatry, 29*(1), 39–47. https://doi.org/10.1097/YCO.0000000000000220

R Core Team. (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing [Computer software] Version v4.3.1. https://www.R-project.org/.

Raes, A. K., De Raedt, R., Verschuere, B., & De Houwer, J. (2009). Failure to loose fear: The impact of cognitive load and trait anxiety on extinction. *Behaviour Research and Therapy, 47*(12), 1096–1101. https://doi.org/10.1016/j.brat.2009.08.002

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory, 21*(6), 64–99. https://doi.org/10.1101/gr.110528.110

Resnik, J., Sobel, N., & Paz, R. (2011). Auditory aversive learning increases discrimination thresholds. *Nature Neuroscience, 14*(6), 791–796. https://doi.org/10.1038/nn.2802

Rowe, M. K., & Craske, M. G. (1998). Effects of varied-stimulus exposure training on fear reduction and return of fear. *Behaviour Research and Therapy, 36*(7–8), 719–734. https://doi.org/10.1016/S0005-7967(97)10017-1

Schechtman, E., Laufer, O., & Paz, R. (2010). Negative valence widens generalization of learning. *Journal of Neuroscience, 30*(31), 10460–10464. https://doi.org/10.1523/JNEUROSCI.2377-10.2010

Scheveneels, S., Boddez, Y., Vervliet, B., & Hermans, D. (2016). The validity of laboratory-based treatment research: Bridging the gap between fear extinction and exposure treatment. *Behaviour Research and Therapy, 86*, 87–94. https://doi.org/10.1016/j.brat.2016.08.015

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): A self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science, 2*(6), Article 140343. https://doi.org/10.1098/rsos.140343

Shiban, Y., Schelhorn, I., Pauli, P., & Mühlberger, A. (2015). Effect of combined multiple contexts and multiple stimuli exposure in spider phobia: A randomized clinical trial in virtual reality. *Behaviour Research and Therapy, 71*, 45–53. https://doi.org/10.1016/j.brat.2015.05.014

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2023). afex: Analysis of factorial experiments Version v1.3.0. http://CRAN.R-project.org/package=afex.

Stantić, M., Brewer, R., Duchaine, B., Banissy, M. J., Bate, S., Susilo, T., Catmur, C., & Bird, G. (2021). The Oxford face matching test: A non-biased test of the full range of individual differences in face perception. *Behavior Research Methods, 54*(1), 158–173. https://doi.org/10.3758/s13428-021-01609-2

Stantić, M., Knyspel, J., Gaches, A., Liu, Y., Bird, G., & Catmur, C. (2024). The Oxford face matching test: Short-form alternative. *Quarterly Journal of Experimental Psychology, 77*(4), 893–897. https://doi.org/10.1177/17470218231182933

Stemerding, L. E., van Ast, V. A., & Kindt, M. (2023). Manipulating expectancy violations to strengthen the efficacy of human fear extinction. *Behaviour Research and Therapy, 165*, Article 104319. https://doi.org/10.1016/j.brat.2023.104319

Struyf, D., Hermans, D., & Vervliet, B. (2018). Maximizing the generalization of fear extinction: Exposures to a peak generalization stimulus. *Behaviour Research and Therapy, 111*, 1–8. https://doi.org/10.1016/j.brat.2018.09.005

Struyf, D., Zaman, J., Hermans, D., & Vervliet, B. (2017). Gradients of fear: How perception influences fear generalization. *Behaviour Research and Therapy, 93*, 116–122. https://doi.org/10.1016/j.brat.2017.04.001

van Dis, E. A. M., van Veen, S. C., Hagenaars, M. A., Batelaan, N. M., Bockting, C. L. H., van den Heuvel, R. M., Cuijpers, P., & Engelhard, I. M. (2020). Long-term outcomes of cognitive behavioral therapy for anxiety-related disorders: A systematic review and meta-analysis. *JAMA Psychiatry, 77*(3), 265–273. https://doi.org/10.1001/jamapsychiatry.2019.3986

Vervliet, B., Craske, M. G., & Hermans, D. (2013). Fear extinction and relapse: State of the art. *Annual Review of Clinical Psychology, 9*(1), 215–248. https://doi.org/10.1146/annurev-clinpsy-050212-185542

Vervliet, B., Vansteenwegen, D., Baeyens, F., Hermans, D., & Eelen, P. (2005). Return of fear in a human differential conditioning paradigm caused by a stimulus change after extinction. *Behaviour Research and Therapy, 43*(3), 357–371. https://doi.org/10.1016/j.brat.2004.02.005

Vervliet, B., Vansteenwegen, D., & Eelen, P. (2004). Generalization of extinguished skin conductance responding in human fear conditioning. *Learning & Memory, 11*(5), 555–558. https://doi.org/10.1101/lm.77404

Vervliet, B., Vansteenwegen, D., & Eelen, P. (2006). Generalization gradients for acquisition and extinction in human contingency learning. *Experimental Psychology, 53*(2), 132–142. https://doi.org/10.1027/1618-3169.53.2.132

Vervoort, E., Vervliet, B., Bennett, M., & Baeyens, F. (2014). Generalization of human fear acquisition and extinction within a novel arbitrary stimulus category. *PLoS One, 9*(5), Article e96569. https://doi.org/10.1371/journal.pone.0096569

Wahl, I., Löwe, B., & Rose, M. (2011). Das Patient-Reported Outcomes Measurement Information System (PROMIS): Übersetzung der Item-Banken für Depressivität und Angst ins Deutsche. *Klinische Diagnostik Und Evaluation, 3*, 236–261.

Waters, A. M., Kershaw, R., & Lipp, O. V. (2018). Multiple fear-related stimuli enhance physiological arousal during extinction and reduce physiological arousal to novel stimuli and the threat conditioned stimulus. *Behaviour Research and Therapy, 106*, 28–36. https://doi.org/10.1016/j.brat.2018.04.005

Waters, A. M., Ryan, K. M., Luck, C. C., Craske, M. G., & Lipp, O. V. (2021). The effects of presenting additional stimuli resembling the CS+ during extinction on extinction retention and generalisation to novel stimuli. *Behaviour Research and Therapy, 144*, Article 103921. https://doi.org/10.1016/j.brat.2021.103921

Waters, A. M., Ryan, K. M., Luck, C. C., Craske, M. G., & Lipp, O. V. (2023). Approximating exposure therapy in the lab: Replacing the CS+ with a similar versus a different stimulus and including additional stimuli resembling the CS+ during extinction. *Behaviour Research and Therapy, 167*, Article 104357. https://doi.org/10.1016/j.brat.2023.104357

Wong, A. H. K., Glück, V. M., Boschet, J. M., & Engelke, P. (2020). Generalization of extinction with a generalization stimulus is determined by learnt threat beliefs. *Behaviour Research and Therapy, 135*, Article 103755. https://doi.org/10.1016/j.brat.2020.103755

Wong, A. H. K., & Lovibond, P. F. (2020). Generalization of extinction of a generalization stimulus in fear learning. *Behaviour Research and Therapy, 125*, Article 103535. https://doi.org/10.1016/j.brat.2019.103535

Zaman, J., Ceulemans, E., Hermans, D., & Beckers, T. (2019). Direct and indirect effects of perception on generalization gradients. *Behaviour Research and Therapy, 114*, 44–50. https://doi.org/10.1016/j.brat.2019.01.006

Zaman, J., Struyf, D., Ceulemans, E., Beckers, T., & Vervliet, B. (2019). Probing the role of perception in fear generalization. *Scientific Reports, 9*(1), 1–8. https://doi.org/10.1038/s41598-019-46176-x

Zaman, J., Vlaeyen, J. W. S., Van Oudenhove, L., Wiech, K., & Van Diest, I. (2015). Associative fear learning and perceptual discrimination: A perceptual pathway in the development of chronic pain. *Neuroscience & Biobehavioral Reviews, 51*, 118–125. https://doi.org/10.1016/j.neubiorev.2015.01.009

Zbozinek, T. D., & Craske, M. G. (2018). Pavlovian extinction of fear with the original conditional stimulus, a generalization stimulus, or multiple generalization stimuli. *Behaviour Research and Therapy, 107*, 64–75. https://doi.org/10.1016/j.brat.2018.05.009